|  |
| --- |
| 3GPP TR 38.843 V18.0.0 (2023-12) |
| Technical Report |
| **3rd Generation Partnership Project;****Technical Specification Group Radio Access Network;**Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface(Release 18) |
|   |
|  |  |
|  |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP.The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented.This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification.Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices. |

|  |
| --- |
|  |
| ***3GPP***Postal address3GPP support office address650 Route des Lucioles - Sophia AntipolisValbonne - FRANCETel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16Internethttp://www.3gpp.org |
| ***Copyright Notification***No part may be reproduced except as authorized by written permission.The copyright and the foregoing restriction extend to reproduction in all media.© 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).All rights reserved.UMTS™ is a Trade Mark of ETSI registered for the benefit of its members3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational PartnersLTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational PartnersGSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 5

Introduction 6

1 Scope 7

2 References 9

3 Definitions of terms, symbols and abbreviations 9

3.1 Terms 9

3.2 Symbols 11

3.3 Abbreviations 11

4 General AI/ML framework 12

4.1 Description of AI/ML stages 12

4.2 Life cycle management 12

4.3 Collaboration levels 15

4.4 Functional framework details 16

5 Use cases 16

5.1 CSI feedback enhancement 16

5.2 Beam management 19

5.3 Positioning accuracy enhancements 21

6 Evaluations 21

6.1 Common evaluation methodology and KPIs 22

6.2 CSI feedback enhancement 23

6.2.1 Evaluation assumptions, methodology and KPIs 23

6.2.2 Performance results 33

6.2.2.1 1-on-1 joint training for CSI compression 35

6.2.2.2 Generalization evaluations for CSI compression 45

6.2.2.3 Scalability evaluations for CSI compression 48

6.2.2.4 Multi-vendor joint training for CSI compression 50

6.2.2.5 Separate training for CSI compression 51

6.2.2.6 Basic performance for CSI prediction 55

6.2.2.7 Generalization evaluations for CSI prediction 58

6.2.2.8 Summary of Performance Results for CSI feedback enhancement 59

6.3 Beam management 59

6.3.1 Evaluation assumptions, methodology and KPIs 59

6.3.2 Performance results 69

6.3.2.1 Basic performance for BM-Case1 69

6.3.2.1.1 Performance when Set B is a subset of Set A for DL Tx beam prediction 70

6.3.2.1.2 Performance when Set B is different than Set A for DL Tx beam prediction 72

6.3.2.1.3 Performance when Set B is a subset of Set A for DL Tx-Rx beam pair prediction 73

6.3.2.1.4 Performance when Set B is different to Set A for DL Tx-Rx beam pair prediction 76

6.3.2.2 Basic performance for BM-Case2 76

6.3.2.2.1 Performance when Set A = Set B 76

6.3.2.2.2 Performance when Set B is a subset of Set A 85

6.3.2.3 Performance under different assumptions/scenarios for BM-Case1 and/or BM-Case2 93

6.3.2.4 Generalization Performance for BM-Case1 and BM-Case2 101

6.3.2.5 Summary of Performance Results for Beam Management 106

6.4 Positioning accuracy enhancements 108

6.4.1 Evaluation assumptions, methodology and KPIs 108

6.4.2 Performance results 115

6.4.2.1 Training Data Collection 116

6.4.2.2 Generalization Aspects 116

6.4.2.3 Fine-tuning 120

6.4.2.4 Model-input Size Reduction 127

6.4.2.5 Non-ideal label(s) 130

6.4.2.6 Summary of Performance Results for Positioning accuracy enhancements 131

7 Potential specification impact assessment 131

7.1 General observations 131

7.2 Physical layer aspects 131

7.2.1 Common framework 131

7.2.2 CSI feedback enhancement 132

7.2.3 Beam management 135

7.2.4 Positioning accuracy enhancements 139

7.3 Protocol aspects 145

7.3.1 Common framework 145

7.3.2 CSI feedback enhancement 145

7.3.3 Beam management 145

7.3.4 Positioning accuracy enhancements 145

7.4 Interoperability and testability aspects 145

7.4.1 Common framework 145

7.4.2 CSI feedback enhancement 145

7.4.3 Beam management 145

7.4.4 Positioning accuracy enhancements 145

8 Conclusions 145

Annex <X> : Change history 146

# Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

In the present document, modal verbs have the following meanings:

**shall** indicates a mandatory requirement to do something

**shall not** indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

**should** indicates a recommendation to do something

**should not** indicates a recommendation not to do something

**may** indicates permission to do something

**need not** indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

**can** indicates that something is possible

**cannot** indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

**will** indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**will not** indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**might** indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

**might not** indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

**is** (or any other verb in the indicative mood) indicates a statement of fact

**is not** (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

# 1 Scope

The application of AI/ML to wireless communications has been thus far limited to implementation-based approaches, both, at the network and the UE sides. A study on enhancement for data collection for NR and ENDC (*FS\_NR\_ENDC\_data\_collect*) has examined the *functional framework for RAN intelligence enabled by further enhancement of data collection through use cases, examples etc. and identify the potential standardization impacts on current NG-RAN nodes and interfaces*. In SA WG2 AI/ML related study, a network functionality NWDAF (Network Data Analytics Function) was introduced in Rel-15 and has been enhanced in Rel-16 and Rel-17.

This study explores the benefits of augmenting the air-interface with features enabling improved support of AI/ML. The 3GPP framework for AI/ML is studied for air-interface corresponding to each target use case regarding aspects such as performance, complexity, and potential specification impact.

Through studying a few carefully selected use cases, assessing their performance in comparison with traditional methods and the associated potential specification impacts that enable their solutions, this study lays the foundation for future air-interface use cases leveraging AI/ML techniques.

Sufficient use cases are targeted to enable the identification of a common AI/ML framework, including functional requirements of AI/ML architecture, which could be used in subsequent projects. The study also serves identifying areas where AI/ML could improve the performance of air-interface functions.

The study serves identifying what is required for an adequate AI/ML model characterization and description establishing pertinent notation for discussions and subsequent evaluations. Various levels of collaboration between the gNB and UE are identified and considered.

Evaluations to exercise the attainable gains of AI/ML based techniques for the use cases under consideration are carried out with the corresponding identification of KPIs with the goal to have a better understanding of the attainable gains and associated complexity requirements.

Finally, specification impact are assessed in order to improve the overall understanding of what would be required to enable AI/ML techniques for the air-interface.

The central objective of this project is to study the 3GPP framework for AI/ML for air-interface corresponding to each target use case regarding aspects such as performance, complexity, and potential specification impact.

The use cases to focus include:

- CSI feedback enhancement

- Spatial-frequency domain CSI compression using two-sided AI model

- Time domain CSI prediction using UE sided model

- Beam management

- Spatial-domain Downlink beam prediction for Set A of beams based on measurement results of Set B of beams

- Temporal Downlink beam prediction for Set A of beams based on the historic measurement results of Set B of beams

- Positioning accuracy enhancements

- Direct AI/ML positioning

- AI/ML assisted positioning

Note: the selection of use cases for this study solely targets the formulation of a framework to apply AI/ML to the air-interface for these and other use cases. The selection itself does not intend to provide any indication of the prospects of any future normative project.

This study also introduces AI/ML model terminology and description to identify common and specific characteristics for framework investigations, namely to:

- Characterize the defining stages of AI/ML related algorithms and associated complexity:

- Model generation, e.g., model training (including input/output, pre-/post-process, online/offline as applicable), model validation, model testing, as applicable

- Inference operation, e.g., input/output, pre-/post-process, as applicable

- Identify various levels of collaboration between UE and gNB pertinent to the selected use cases, e.g.,

- No collaboration: implementation-based only AI/ML algorithms without information exchange [for comparison purposes]

- Various levels of UE/gNB collaboration targeting at separate or joint ML operation.

- Characterize lifecycle management of AI/ML model: e.g., model training, model deployment, model inference, model monitoring, model updating

- Dataset(s) for training, validation, testing, and inference

- Identify common notation and terminology for AI/ML related functions, procedures and interfaces

- Note: the work done for *FS\_NR\_ENDC\_data\_collect* is considered when appropriate

For the use cases under consideration:

1) Performance benefits of AI/ML based algorithms for the agreed use cases are evaluated:

- Methodology based on statistical models (from TR 38.901 and TR 38.857 [positioning]), for link and system level simulations.

- Extensions of 3GPP evaluation methodology for better suitability to AI/ML based techniques should be considered as needed.

- Whether field data are optionally needed to further assess the performance and robustness in real-world environments should be discussed as part of the study.

- Need for common assumptions in dataset construction for training, validation and test for the selected use cases.

- Consider adequate model training strategy, collaboration levels and associated implications

- Consider agreed-upon base AI model(s) for calibration

- AI model description and training methodology used for evaluation should be reported for information and cross-checking purposes

- KPIs: Determine the common KPIs and corresponding requirements for the AI/ML operations. Determine the use-case specific KPIs and benchmarks of the selected use-cases.

- Performance, inference latency and computational complexity of AI/ML based algorithms should be compared to that of a state-of-the-art baseline

- Overhead, power consumption (including computational), memory storage, and hardware requirements (including for given processing delays) associated with enabling respective AI/ML scheme, as well as generalization capability should be considered.

2) Potential specification impact, specifically for the agreed use cases and for a common framework, is assessed:

- PHY layer aspects, e.g., (RAN1)

- Considering aspects related to, e.g., the potential specification of the AI Model lifecycle management, and dataset construction for training, validation and test for the selected use cases

- Use case and collaboration level specific specification impact, such as new signalling, means for training and validation data assistance, assistance information, measurement, and feedback

- Protocol aspects, e.g., (RAN2) - RAN2 only starts the work after there is sufficient progress on the use case study in RAN1

- Considering aspects related to, e.g., capability indication, configuration and control procedures (training/inference), and management of data and AI/ML model, per RAN1 input

- Collaboration level specific specification impact per use case

- Interoperability and testability aspects, e.g., (RAN4) - RAN4 only starts the work after there is sufficient progress on use case study in RAN1 and RAN2

- Requirements and testing frameworks to validate AI/ML based performance enhancements and ensuring that UE and gNB with AI/ML meet or exceed the existing minimum requirements if applicable

- Considering the need and implications for AI/ML processing capabilities definition

Note 1: Specific AI/ML models are not expected to be specified and are left to implementation. User data privacy needs to be preserved.

Note 2: The study on AI/ML for air interface is based on the current RAN architecture and new interfaces shall not be introduced.

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] RP-213599: "New SI: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface", Qualcomm (Moderator).

[3] 3GPP TR 38.901: "Study on channel model for frequencies from 0.5 to 100 GHz".

[4] 3GPP TR 38.857: "Study on NR positioning enhancements".

[5] 3GPP TR 38.802: "Study on new radio access technology Physical layer aspects".

# 3 Definitions of terms, symbols and abbreviations

## 3.1 Terms

For the purposes of the present document, the terms given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

**AI/ML-enabled Feature:** refers to a Feature where AI/ML may be used.

**AI/ML Model:** A data driven algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs.

**AI/ML model delivery:** A generic term referring to delivery of an AI/ML model from one entity to another entity in any manner. Note: An entity could mean a network node/function (e.g., gNB, LMF, etc.), UE, proprietary server, etc.

**AI/ML model Inference:**  A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs.

**AI/ML model testing:** A subprocess of training, to evaluate the performance of a final AI/ML model using a dataset different from one used for model training and validation. Differently from AI/ML model validation, testing does not assume subsequent tuning of the model.

**AI/ML model training:** A process to train an AI/ML Model [by learning the input/output relationship] in a data driven manner and obtain the trained AI/ML Model for inference.

**AI/ML model transfer:** Delivery of an AI/ML model over the air interface in a manner that is not transparent to 3GPP signalling, either parameters of a model structure known at the receiving end or a new model with parameters. Delivery may contain a full model or a partial model.

**AI/ML model validation:** A subprocess of training, to evaluate the quality of an AI/ML model using a dataset different from one used for model training, that helps selecting model parameters that generalize beyond the dataset used for model training.

**Data collection:** A process of collecting data by the network nodes, management entity, or UE for the purpose of AI/ML model training, data analytics and inference.

**Federated learning / federated training:** A machine learning technique that trains an AI/ML model across multiple decentralized edge nodes (e.g., UEs, gNBs) each performing local model training using local data samples. The technique requires multiple interactions of the model, but no exchange of local data samples.

**Functionality identification:** A process/method of identifying an AI/ML functionality for the common understanding between the NW and the UE. Note: Information regarding the AI/ML functionality may be shared during functionality identification. Where AI/ML functionality resides depends on the specific use cases and sub use cases.

**Management instruction:** Information needed to ensure proper inference operation. This information may include selection/(de)activation/switching of AI/ML models or AI/ML functionalities, fallback to non-AI/ML operation, etc.

**Model activation:** enable an AI/ML model for a specific AI/ML-enabled feature.

**Model deactivation:** disable an AI/ML model for a specific AI/ML-enabled feature.

**Model download:** Model transfer from the network to UE.

**Model identification:** A process/method of identifying an AI/ML model for the common understanding between the NW and the UE. Note: The process/method of model identification may or may not be applicable. Note: Information regarding the AI/ML model may be shared during model identification.

**Model monitoring:** A procedure that monitors the inference performance of the AI/ML model.

**Model parameter update:** Process of updating the model parameters of a model.

**Model selection:** The process of selecting an AI/ML model for activation among multiple models for the same AI/ML enabled feature. Note: Model selection may or may not be carried out simultaneously with model activation.

**Model switching:** Deactivating a currently active AI/ML model and activating a different AI/ML model for a specific AI/ML-enabled feature.

**Model update:** Process of updating the model parameters and/or model structure of a model.

**Model upload:** Model transfer from UE to the network.

**Network-side (AI/ML) model:** An AI/ML Model whose inference is performed entirely at the network.

**Offline field data:** The data collected from field and used for offline training of the AI/ML model.

**Offline training:** An AI/ML training process where the model is trained based on collected dataset, and where the trained model is later used or delivered for inference. Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as offline training by commonly accepted conventions.

**Online field data:** The data collected from field and used for online training of the AI/ML model.

**Online training:** An AI/ML training process where the model being used for inference) is (typically continuously) trained in (near) real-time with the arrival of new training samples. Note: the notion of (near) real-time vs. non real-time is context-dependent and is relative to the inference time-scale. Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as online training by commonly accepted conventions. Note: Fine-tuning/re-training may be done via online or offline training.

**Reinforcement Learning (RL):** A process of training an AI/ML model from input (a.k.a. state) and a feedback signal (a.k.a. reward) resulting from the model’s output (a.k.a. action) in an environment the model is interacting with.

**Semi-supervised learning:** A process of training a model with a mix of labelled data and unlabelled data.

**Supervised learning:** A process of training a model from input and its corresponding *labels*.

**Test encoder/decoder for TE:** AI/ML model for UE encoder/gNB decoder implemented by TE.

**Two-sided (AI/ML) model:** A paired AI/ML Model(s) over which joint inference is performed, where joint inference comprises AI/ML Inference whose inference is performed jointly across the UE and the network, i.e, the first part of inference is firstly performed by UE and then the remaining part is performed by gNB, or vice versa.

**UE-side (AI/ML) model:** An AI/ML Model whose inference is performed entirely at the UE.

**Unsupervised learning:** A process of training a model without labelled data.

**Proprietary-format models**: ML models of vendor-/device-specific proprietary format, from 3GPP perspective. They are not mutually recognizable across vendors and hide model design information from other vendors when shared. Note: An example is a device-specific binary executable format.

**Open-format models**: ML models of specified format that are mutually recognizable across vendors and allow interoperability, from 3GPP perspective. They are mutually recognizable between vendors and do not hide model design information from other vendors when shared.

## 3.2 Symbols

For the purposes of the present document, the following symbols apply:

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

AI Artificial Intelligence

API Application Programming Interface

BM Beam Management

CDF Cumulative Distribution Function

CIR Channel Impulse Response

CNN Convolutional Neural Network

CQI Channel Quality Indicator

CSI Channel State Information

DL Downlink

DP Delay profile

EVM Evaluation Methodology

FLOP Floating point Operation\*

GCS Generalized Cosine Similarity

KPI Key Performance Indicator

L1-RSRP Layer 1 reference signal received power

LCM Life Cycle Management

LLS Link Level Simulations

LMF Location Management Function

LOS Line-of-Sight

LPP LTE Positioning Protocol

ML Machine Learning

NLOS Non-Line-of-Sight

NMSE Normalized Mean Square Error

NRPPa NR Positioning Protocol A

NW Network

PDP Power Delay Profile

PRS Positioning Reference Signal

PRU Positioning Reference Unit

RNN Recurrent Neural Network

RI Rank Indicator

RU Resource Utilization

SGCS Squared Generalized Cosine Similarity

SLS System Level Simulations

SQ Scalar Quantization

SRS Sounding Reference Signal

TRP Transmission-Reception Point

TxRU Transceiver Unit

UE User Equipment

UPT User Perceived Throughput

VQ Vector Quantization

\* Note: Caution is advised not to confuse FLOPs (the plural form of FLOP), which a measure of model complexity, with FLOPS (FLoating point OPerations per Second), which is a measure of compute performance of a device.

# 4 General AI/ML framework

The purpose of this clause is to identify common notation and terminology for AI/ML related functions, procedures and interfaces.

Note: The work done for FS\_NR\_ENDC\_data\_collect is considered when appropriate.

## 4.1 Description of AI/ML stages

In this clause, the defining stages of AI/ML related algorithms and associated complexity are characterized, namely:

- Model generation, e.g., model training (including input/output, pre-/post-process, online/offline as applicable), model validation, model testing, as applicable

- Inference operation, e.g., input/output, pre-/post-process, as applicable

In addition, the treatment of dataset(s) for training, validation, testing, and inference is documented.

## 4.2 Life Cycle Management

In this clause, the life cycle management (LCM) of AI/ML model (e.g., model training, model deployment, model inference, model monitoring, model updating) and AI/ML functionality are characterized.

The following aspects, including the definition of components (if needed) and necessity, are studied in LCM:

- Data collection

- Note: This also includes associated assistance information, if applicable.

- Model training

- Functionality/model identification

- Model delivery/transfer

- Model inference operation

- Functionality/model selection, activation, deactivation, switching, and fallback operation.

- Including: Decision by the network (either network initiated or UE-initiated and requested to the network), decision by the UE (event-triggered as configured by the network, UE’s decision reported to the network, or UE-autonomous either with UE’s decision reported to the network or without it)

- Functionality/model monitoring

- Model update

- UE capability

Note: Some aspects in the list may not have specification impact.

### 4.2.1 LCM Flavours

The LCM procedure is studied for the case that an AI/ML model has a *model ID* with associated information and/or for the case that a given *functionality* is provided by some AI/ML operations. Note: Applicability of functionality-based LCM and model-ID-based LCM is a separate discussion.

From RAN1 perspective, an AI/ML model identified by a model ID may be *logical*, and how it maps to physical AI/ML model(s) may be up to implementation. When distinction is necessary for discussion purposes, companies may use the term a *logical AI/ML model* to refer to a model that is identified and assigned a model ID, and *physical AI/ML model(s)* to refer to an actual implementation of such a model.

For UE-side models and UE-part of two-sided models:

- For AI/ML functionality identification

- Legacy 3GPP framework of feature is taken as a starting point.

- UE indicates supported functionalities/functionality for a given sub-use-case.

- UE capability reporting is taken as starting point.

- For AI/ML model identification

- Models are identified by model ID at the Network. UE indicates supported AI/ML models.

In *functionality-based* LCM, network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signalling (e.g., RRC, MAC-CE, DCI). Models may not be identified at the Network, and UE may perform model-level LCM. Whether and how much awareness/interaction NW should have about model-level LCM requires further study. For functionality identification, there may be either one or more than one Functionalities defined within an AI/ML-enabled feature, whereby AI/ML-enabled Feature refers to a Feature where AI/ML may be used. Note: UE may have one AI/ML model for the functionality, or UE may have multiple AI/ML models for the functionality.

For *AI/ML functionality identification* and *functionality-based LCM* of UE-side models and/or UE-part of two-sided models, *functionality* refers to an AI/ML-enabled Feature/FG enabled by configuration(s), where configuration(s) is(are) supported based on conditions indicated by UE capability. Correspondingly, *functionality-based LCM* operates based on, at least, one configuration of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature/FG.

After *functionality identification*, necessity, mechanisms, for UE to report updates on applicable functionality(es) among functionality(es) are studied, where the applicable functionalities may be a subset of all functionalities. Applicable functionalities can be reported by the UE.

In *model-ID-based* LCM, models are identified at the Network, and Network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.

For *AI/ML model identification* and *model-ID-based LCM* of UE-side models and/or UE-part of two-sided models, *model-ID-based LCM* operates based on identified models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled Feature/FG and additional conditions (e.g., scenarios, sites, and datasets) as determined/identified between UE-side and NW-side.

After model identification, necessity, mechanisms, for UE to report updates on applicable UE part/UE-side model(s), are studied, where the applicable models may be a subset of all identified models. Applicable models can be reported by the UE.

How to handle the impact of UE’s internal conditions such as memory, battery, and other hardware limitations on functionality/model operations and AI/ML-enabled Feature is to be studied. Note: it does not preclude any existing solutions.

For functionality/model-ID based LCM, once functionalities/models are identified, the same or similar procedures may be used for their activation, deactivation, switching, fallback, and monitoring.

Model ID, if needed, can be used in a Functionality (defined in functionality-based LCM) for LCM operations.

### 4.2.2 Model identification

For *AI/ML model identification* of UE-side or UE-part of two-sided models, model identification is categorized in the following types:

- Type A: Model is identified to NW (if applicable) and UE (if applicable) without over-the-air signalling

- The model may be assigned with a model ID during the model identification, which may be referred/used in over-the-air signalling after model identification.

- Type B: Model is identified via over-the-air signalling,

- Type B1:

- Model identification initiated by the UE, and NW assists the remaining steps (if any) of the model identification

- the model may be assigned with a model ID during the model identification

- Type B2:

- Model identification initiated by the NW, and UE responds (if applicable) for the remaining steps (if any) of the model identification

- the model may be assigned with a model ID during the model identification

- Note: This study does not imply that model identification is necessary.

One example use case for Type B1 and B2 is model identification in model transfer from NW to UE. Another example is model identification with data collection related configuration(s) and/or indication(s) and/or dataset transfer. Note: Other example use cases are not precluded. Note: Offline model identification may be applicable for some of the example use cases.

Once models are identified, at least for Type A, UE can indicate supported AI/ML model IDs for a given AI/ML-enabled Feature/FG in a UE capability report as starting point. Note: model identification using capability report is not precluded for type B1 and type B2.

Model ID may or may not be globally unique, and different types of model IDs may be created for a single model for various LCM purposes. Note: Details can be studied in the WI phase.

### 4.2.3 Additional conditions

For an AI/ML-enabled feature/FG, *additional conditions* refer to any aspects that are assumed for the training of the model but are not a part of UE capability for the AI/ML-enabled feature/FG. It does not imply that *additional conditions* are necessarily specified. *Additional conditions* can be divided into two categories: NW-side additional conditions and UE-side additional conditions. Note: whether specification impact is needed is a separate discussion.

For inference for UE-side models, to ensure consistency between training and inference regarding NW-side *additional conditions* (if identified), the following options can be taken as potential approaches (when feasible and necessary):

- Model identification to achieve alignment on the NW-side additional condition between NW-side and UE-side

- Model training at NW and transfer to UE, where the model has been trained under the additional condition

- Information and/or indication on NW-side additional conditions is provided to UE

- Consistency assisted by monitoring (by UE and/or NW, the performance of UE-side candidate models/functionalities to select a model/functionality)

- Other approaches are not precluded

- Note: the possibility that different approaches can achieve the same function is not denied

### 4.2.4 Scenario/configuration specific Models

Scenario/configuration specific (including site-specific configuration/channel conditions) models may provide performance benefits in some studied use cases (i.e., when a single model cannot generalize well to multiple scenarios/configurations/sites).

- At least, when UE has limitation to store all related models, model delivery/transfer, if feasible, to UE may be beneficial, at the cost of overhead/latency associated with model delivery/transfer.

- Note: On-device Finetuning/retraining, if feasible, of a single model may be an alternative to model delivery/transfer.

- Note: a single model may generalize well in some studied use cases.

- Note: Model transfer/delivery to UE may also face challenges, e.g., proprietary issues /burdens in some scenarios

Various approaches for achieving good performance across different scenarios/configurations/sites are studied, including

*- Model generalization*, i.e., using one model that is generalizable to different scenarios/configurations/sites

*- Model switching*, i.e., switching among a group of models where each model is for a particular scenario/configuration/site

- Models in a group of models may have varying model structures, share a common model structure, or partially share a common sub-structure. Models in a group of models may have different input/output format and/or different pre-/post-processing.

*- Model update*, i.e., using one model whose parameters are flexibly updated as the scenario/configuration/site that the device experiences changes over time. Fine-tuning is one example.

### 4.2.5 Data collection

Data collection may be performed for different purposes in LCM, e.g., model training, model inference, model monitoring, model selection, model update, etc. each may be done with different requirements and potential specification impact.

For all types of offline model training (i.e., UE- /NW-/ two-sided model training), there is no latency requirement for data collection. For model inference, when required data comes from other entities, there is a latency requirement for data collection. For (real-time) performance monitoring, when required monitoring data (e.g., performance metric) comes from other entities, there is a latency requirement for data collection.

At least for the use cases studied in this study item, it is assumed that the analysis/selection of the data collection frameworks should focus on the RRC\_CONNECTED state (for both data generation and reporting). Analysis and potential enhancement of the non-connected state can be revisited when needed. Note that existing specification supports DL PRS measurement and UE positioning in both RRC\_CONNECTED and RRC\_INACTIVE state.

At least the following aspects, if applicable, are considered along with the corresponding specification impact:

- Measurement configuration and reporting

- Contents, type and format of data including:

- Data related to model input

- Data related to ground-truth

- Quality of the data

- Other information

- Signalling of assistance information for categorizing the data

- Note: The study should consider the feasibility of disclosure of proprietary information

- Signalling for data collection procedure

## 4.3 Collaboration levels

In this clause, various levels of collaboration between UE and Network are identified as found pertinent to the selected use cases, e.g.,

- No collaboration: implementation-based only AI/ML algorithms without information exchange for comparison purposes

- Various levels of UE/Network collaboration targeting at separate or joint ML operation

The following Network-UE collaboration levels are considered as one aspect for defining collaboration levels

1. **Level x**: No collaboration.

2. **Level y**: Signalling-based collaboration without model transfer. Note: this level includes cases without model delivery.

3. **Level z**: Signalling-based collaboration with model transfer.

Level x/y boundary is understood such as Level x is implementation-based AI/ML operation without any dedicated AI/ML-specific enhancement (e.g., LCM related signalling, RS) collaboration between network and UE. (Note: The AI/ML operation may rely on future specification not related to AI/ML collaboration. The AI/ML approaches can be used as baseline for performance evaluation for future releases.)

Level y/z boundary is defined based on whether model delivery over the air interface is done in a non-transparent manner to 3GPP signalling. Note: procedures other than model transfer/delivery are decoupled with collaboration Level y-z.

Table 4.3-1 introduces different options for model delivery/transfer to UE, training location, and model delivery/transfer format combinations for UE-side models and UE-part of two-sided models:

Table 4.3-1: Model delivery/transfer cases

|  |  |  |  |
| --- | --- | --- | --- |
| **Case** | **Model delivery/transfer** | **Model storage location** | **Training location** |
| **y** | model delivery (if needed) over-the-top. | Outside 3GPP Network | UE-side / NW-side / neutral site |
| **z1** | model transfer in proprietary format. | 3GPP Network | UE-side / neutral site |
| **z2** | model transfer in proprietary format. | 3GPP Network | NW-side |
| **z3** | model transfer in open format. | 3GPP Network | UE-side / neutral site |
| **z4** | model transfer in open format of a *known model structure* at UE, i.e., an exact model structure as has been previously identified between NW and UE and for which the UE has explicitly indicated its support.  | 3GPP Network | NW-side |
| **z5** | model transfer in open format of *an unknown model structure* at UE, i.e., any other model structure not covered in z4, including any model structure that is only partially known. | 3GPP Network | NW-side |
| Note: The definition of various Cases is only for the purpose of facilitating discussion and does not imply applicability, feasibility, entity mapping, architecture, signalling nor any prioritization. |

When a model of a known structure at UE (e.g., Case z4) is transferred from the Network, the new model being identified (e.g., via Type B2) has the same structure as a previously identified model at the Network and UE.

For model delivery/transfer to UE (for UE-side models and UE-part of two-sided models):

- Model delivery/transfer to UE, if feasible, may be beneficial to handle scenario/configuration specific (including site-specific configuration/channel conditions) models (i.e., when a single model cannot generalize well to multiple scenarios/configurations/sites), to reduce the device storage requirement.

- Model delivery/transfer to UE after offline compiling and/or testing may be friendlier from UE’s implementation point of view compared to the case without offline compiling and/or testing. On the other hand, the case without offline compiling and/or testing (that can update parameter with known model structure), may have benefit at least in terms of shorter model parameter update timescale.

- Model transfer/delivery of an unknown structure at UE has more challenges related to feasibility (e.g. UE implementation feasibility) compared to delivery/transfer of a known structure at UE.

- For model trained at network side, Case y (w/ NW-side training) and Case z2 may incur the burden of offline cross-vendor collaboration such as sending a model to the UE-side and/or compiling a model.

- For model trained at UE side/neutral site, Case z1 and Case z3 may incur the burden of offline cross-vendor collaboration to send the trained model from the UE-side to the network, compared to Case y (w/ UE-side training) which does not have such burden.

- Model storage at the 3GPP network, compared to storing the model outside the 3GPP network, may come with 3GPP network side burden on model maintenance/storage.

- Proprietary design disclosure concern may arise from model training and/or model storage at the network side compared to other cases (such as case y with UE side training) which does not have such issue.

## 4.4 Functional framework details

This section introduces the functional framework for AI/ML for NR air interface illustrated in Figure 4.4-1. The aim of this framework is to cover a general functional architecture addressing both model-ID-based LCM and functionality-based LCM, introduced in clause 4.2. Therefore, some of the functions or data/information/instruction flows (i.e., the arrows) shown in the Figure 4.4-1 might not always be relevant for a given LCM approach. As an illustrative example, consider a scenario where the network performs functionality-based LCM and where models are not identified in the network, while the UE concurrently performs model-level management (e.g., model selection/switching/(de)activation, etc…). In this hypothetical case, the "Model Training" or "Model Storage" functions with their respective procedures, may be regarded as irrelevant from the network’s perspective.

In clause 7, the functions and data/information/instruction flows (i.e., the arrows) depicted in Figure 4.4-1 are analysed for any standardization impact and its implications.

Note: The functional framework and high-level procedures defined in this TR should not prevent from "thinking beyond" them during a normative phase if any use case requires so.



Figure 4.4-1: Functional framework for AI/ML for NR Air Interface

As seen in Figure 4.4-1, the general framework consists of the following:

- Data Collection is a function that provides input data to the Model Training, Management, and Inference functions.

- Training Data: Data needed as input for the AI/ML Model Training function.

- Monitoring Data: Data needed as input for the Management of AI/ML models or AI/ML functionalities.

- Inference Data: Data needed as input for the AI/ML Inference function.

- Model Training is a function that performs AI/ML model training, validation, and testing which may generate model performance metrics which can be used as part of the model testing procedure. The Model Training function is also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) based on Training Data delivered by a Data Collection function, if required.

- Trained/Updated Model: In case of having a Model Storage function, this is used to deliver trained, validated, and tested AI/ML models to the Model Storage function, or to deliver an updated version of a model to the Model Storage function.

- Management is a function that oversees the operation (e.g., selection/(de)activation/switching/fallback) and monitoring (e.g., performance) of AI/ML models or AI/ML functionalities. This function is also responsible for making decisions to ensure the proper inference operation based on data received from the Data Collection function and the Inference function.

- Management Instruction: Information needed as input to manage the Inference function. Concerning information may include selection/(de)activation/switching of AI/ML models or AI/ML-based functionalities, fallback to non-AI/ML operation (i.e., not relying on inference process), etc…

- Model Transfer/Delivery Request: Used to request model(s) to the Model Storage function.

- Performance Feedback / Retraining Request: Information needed as input for the Model Training function, e.g., for model (re)training or updating purposes.

- Inference is a function that provides outputs from the process of applying AI/ML models or AI/ML functionalities, using the data that is provided by the Data Collection function (i.e., Inference Data) as an input. The Inference function is also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) based on Inference Data delivered by a Data Collection function, if required.

- Inference Output: Data used by the Management function to monitor the performance of AI/ML models or AI/ML functionalities.

- Model Storage is a function responsible for storing trained/updated models that can be used to perform the Inference function.

- Note: The Model Storage function in Figure 4.4-1 is only intended as a reference point (if any) when applicable for protocol terminations, model transfer/delivery, and related processes. It should be stressed that its purpose does not encompass restricting the actual storage locations of models. Therefore, the specification impact of all data/information/instruction flows (i.e., the arrows in Figure 4.4-1) to/from this function should be studied case by case.

- Model Transfer/Delivery: Used to deliver an AI/ML model to the Inference function.

# 5 Use cases

Initial set of use cases includes:

- CSI feedback enhancement, e.g., overhead reduction, improved accuracy, prediction [RAN1]

- Beam management, e.g., beam prediction in time, and/or spatial domain for overhead and latency reduction, beam selection accuracy improvement [RAN1]

- Positioning accuracy enhancements for different scenarios including, e.g., those with heavy NLOS conditions [RAN1]

- The AI/ML approaches for the selected sub use cases need to be diverse enough to support various requirements on the gNB-UE collaboration levels

Note: the selection of use cases for this study solely targets the formulation of a framework to apply AI/ML to the air-interface for these and other use cases. The selection itself does not intend to provide any indication of the prospects of any future normative project.

## 5.1 CSI feedback enhancement

***Finalization of representative sub-use cases*:**

The following are selected as representative sub-use cases:

- Spatial-frequency domain CSI compression using two-sided AI model. Note: All pre-processing/post-processing, quantization/de-quantization are within the scope of the sub use case.

- The study of AI/ML based CSI compression should be based on the legacy CSI feedback signalling framework.

- Time domain CSI prediction using UE-side model.

For CSI compression using two-sided model use case, considered AI/ML model training collaborations include:

- Type 1: Joint training of the two-sided model at a single side/entity, e.g., UE-sided or Network-sided.

- Type 2: Joint training of the two-sided model at network side and UE side, respectively.

- Type 3: Separate training at network side and UE side, where the UE-side CSI generation part and the NW-side CSI reconstruction part are trained by UE side and network side, respectively.

- Note: Joint training means the generation model and reconstruction model should be trained in the same loop for forward propagation and backward propagation. Joint training could be done both at single node or across multiple nodes (e.g., through gradient exchange between nodes).

- Note: Separate training includes sequential training starting with UE side training, or sequential training starting with NW side training

- Note: training collaboration Type 2 over the air interface for model training (not including model update) is concluded to be deprioritized in Rel-18 SI.

For Type 2 (Joint training of the two-sided model at network side and UE side, respectively), note that joint training includes both simultaneous training and sequential training, in which the pros and cons could be discussed separately. Further, note that Type 2 sequential training starts with NW side training.

In CSI compression using two-sided model use case, for discussion of training collaboration Type 1, separate columns are shown for both known model structure, and unknown model structure separately for NW-sided and UE-sided, respectively. Table 5.1-1 captures the pros/cons of training collaboration Type 1 for CSI compression using two-sided model use case.

Table 5.1-1: Pros and Cons of training collaboration Type 1

|  |  |  |
| --- | --- | --- |
| Characteristics \ Training Types | Type 1: NW side | Type 1: UE side |
| Unknown model structure at UE | Known model structure at UE | Unknown model structure at UE | Known model structure at UE |
| Whether model can be kept proprietary  | No | No | No | No |
| Whether require privacy-sensitive dataset sharing | No (Note 1) | No (Note 1) | No (Note 1) | No (Note 1) |
| Flexibility to support cell/site/scenario/configuration specific model | Flexible except for UE defined scenarios. Not flexible for UE defined scenarios unless UE assistance information is supported and available. (Note 6) | Flexible except for UE defined scenarios. Not flexible for UE defined scenarios unless UE assistance information is supported and available.(Note 6) | Flexible except for NW defined scenarios. Not flexible for NW defined scenarios unless NW assistance information is supported and available.(Note 6) | Flexible except for NW defined scenarios. Not flexible for NW defined scenarios unless NW assistance information is supported and available.(Note 6) |
| Whether gNB/device specific optimization is allowed | gNB: YesUE: No | gNB: YesUE: less flexible compared to UE side | gNB: NoUE: Yes | gNB: less flexible compared to NW sideUE: Yes |
| Model update flexibility after deployment  | Flexible only if UE supports the new structure | Flexible for parameter update | Flexible, less flexible than Type 1 NW side | Flexible for parameter update, less flexible than Type 1 NW side |
| Feasibility of allowing UE side and NW side to develop/update models separately | gNB: FeasibleUE: Not feasible due to Type 1 definition.  | gNB: Feasible with restriction for CSI reconstruction model. UE: Not feasible due to Type 1 definition.  | gNB: Not feasible due to Type 1 definition. UE: Feasible. | gNB: Not feasible due to Type 1 definition. UE: Feasible with restriction for CSI generation model.  |
| Whether gNB can maintain/store a single/unified CSI reconstruction model over different UEs (Note 2) | Yes | Yes. Performance refers to observations in "1 NW part model to M>1 UE part models" of clause 6.2.2.4 (Note 4) | No | No |
| Whether UE device can maintain/store a single/unified CSI generation model over different NW vendors (Note 3) | No | No | Yes | Yes. Performance refers to observations in "1 UE part model to N>1 NW part models" of clause 6.2.2.4 (Note 4) |
| Extendibility: to train new UE-side model compatible with NW-side model in use; (Note 5) | Yes | Yes | No consensus | No consensus |
| Extendibility: To train new NW-side model compatible with UE-side model in use; (Note 5) | No consensus | No consensus | Yes | Yes |
| Whether training data distribution can match the inference device | Limited | Limited | Yes | Yes |
| Software/hardware compatibility (Whether device capability can be considered for model development) | No for UE | Yes | No for NW | Yes |
| Model performance  | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 |
| Note 1: Assume precoding matrix is not privacy sensitive data. FFS: other information such as channel matrix and assisted information.Note 2: Whether gNB/UE needs to maintain/store multiple CSI generation/reconstruction models respectively, is not discussed.Note 3: For model inference, UE does not need to use multiple models from different NW vendors per cell.Note 4: One to many joint trainings is assumed.Note 5: The performance of the new model is similar to the performance of sequential training when training Type 1 support freezing a part of two sided model.Note 6: For this table, NW defined scenarios are scenarios with NW defined dataset categorization. UE defined scenarios are scenarios with UE defined dataset categorization. |

Table 5.1-2 captures the pros/cons of training collaboration Type 2 and Type 3 for CSI compression using two-sided model use case.

Table 5.1-2: Pros and Cons of training collaboration Type 2 and Type 3

|  |  |  |
| --- | --- | --- |
| Characteristics \ Training Types | Type 2 | Type 3 |
| Simultaneous | **Sequential**  | NW first |  UE first |
| Whether model can be kept proprietary  | Yes (Note 1) | Yes (Note 1) | Yes (Note 1) | Yes (Note 1) |
| Whether requires privacy-sensitive dataset sharing | No (Note 2) | No (Note 2) | No (Note 2) | No (Note 2) |
| Flexibility to support cell/site/scenario/configuration specific model | No consensus | No consensus | [Semi] flexible except for UE defined scenarios. (Note 3) [Semi] flexible for UE defined scenarios if UE assistance information is supported and available.  | [Semi] flexible except for NW defined scenarios (Note 3). [Semi] flexible for NW defined scenarios if NW assistance information is supported and available.  |
| Whether gNB/device specific optimization is allowed | Yes | Yes | Yes | Yes |
| Model update flexibility after deployment (Note 4) | Not flexible | No consensus | Semi-flexible | Semi-flexible |
| Feasibility of allowing UE side and NW side to develop/update models separately | Infeasible | No consensus | Feasible | Feasible |
| Whether gNB can maintain/store a single/unified model over different UE vendors (Note 5) | Yes. Performance refers to observations in "1 NW part model to M>1 UE part models" and "1 UE part model to N>1 NW part models" of clause 6.2.2.4. | Yes. Performance refers to observations in "1 NW part model to M>1 UE part models" of clause 6.2.2.4. | Yes. Performance refers to observations in "NW first training, 1 NW part model to 1 UE part model, same backbone" and "NW first training, 1 NW part model to 1 UE part model, different backbones" of clause 6.2.2.5. | Yes. Performance refers to observations in "UE first training, M>1 UE part models to 1 NW part model" of clause 6.2.2.5. |
| Whether UE device can maintain/store a single/unified CSI generation model over different NW vendors (Note 6) | Yes. Performance refers to observations in "1 NW part model to M>1 UE part models" and "1 UE part model to N>1 NW part models" of clause 6.2.2.4. | Yes. Performance refers to observations in "1 NW part model to M>1 UE part models" of clause 6.2.2.4. | Yes. Performance refers to observations in "NW first training, 1 UE part model to N>1 NW part models" of clause 6.2.2.5. | Yes. Performance refers to observations in "UE first training, 1 NW part model to 1 UE part model, same backbone". And "UE first training, 1 NW part model to 1 UE part model, different backbones" of clause 6.2.2.5. |
| Extendibility: to train new UE-side model compatible with NW-side model in use;  | Not support | Support | Support | No consensus |
| Extendibility: To train new NW-side model compatible with UE-side model in use | Not support | Not support | No consensus | Support |
| Whether training data distribution can match the inference device | No consensus | Yes for UE-part model, limited for NW-part model | Limited | Yes |
| Software/hardware compatibility (Whether device capability can be considered for model development) | Compatible | Compatible | Compatible | Compatible |
| Model performance | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 | Performance refers to clause 6.2.2 |
| Note 1: Assume information on model structure disclosed in training collaboration does not reveal proprietary information.Note 2: Assume precoding matrix is not privacy sensitive data. FFS: other information such as channel matrix and assisted information.Note 3: For this table, NW defined scenarios are scenarios with NW defined dataset categorization. UE defined scenarios are scenarios with UE defined dataset categorization. [Semi] means no consensus for including "semi".Note 4: Flexibility after deployment is evaluated by the amount of offline cross-vendor co-engineering effort. Flexible indicates minimum additional co-engineering between vendors, semi-flexible indicates additional co-engineering effort between vendors.Note 5: Whether gNB/UE needs to maintain/store multiple CSI generation/reconstruction models respectively, is not discussed.Note 6: For model inference, UE does not need to use multiple models from different NW vendors per cell.  |

For CSI compression use case:

- For *model training*, training data can be generated by UE/gNB

- For NW-part of two-sided *model inference*, input data can be generated by UE and terminated at gNB.

- For UE-part of two-sided *model inference*, input data is internally available at UE.

- For *performance monitoring* at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB

In CSI compression using two-sided model use case, in order to select a CSI generation model compatible with the CSI reconstruction model used by the gNB, the following aspect has been proposed:

- Pairing information can be established based on model identification

For CSI prediction use cases:

- For *model training*, training data can be generated by UE.

- For UE-side *model inference*, input data is internally available at UE.

- For *performance monitoring* at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB.

## 5.2 Beam management

***Finalization of representative sub-use cases*:**

The following are selected as representative sub-use cases:

- BM-Case1: Spatial-domain Downlink beam prediction for Set A of beams based on measurement results of Set B of beams

- Consider: Alt. 1): AI/ML model training and inference at NW side. Alt. 2): AI/ML model training and inference at UE side.

- Consider: Alt. i): Set A and Set B are different (Set B is NOT a subset of Set A). Alt. ii): Set B is a subset of Set A. Note: Set A is for DL beam prediction. The codebook construction of Set A and Set B can be clarified by companies.

- AI/ML model input consider: Alt 1): Only L1-RSRP measurement based on Set B; Alt.2): L1-RSRP measurement based on Set B and assistance information; Alt. 3): CIR based on Set B; Alt. 4): L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID.

- BM-Case2: Temporal Downlink beam prediction for Set A of beams based on the historic measurement results of Set B of beams

- Consider: Alt. 1): AI/ML model training and inference at NW side. Alt. 2): AI/ML model training and inference at UE side.

- Consider: Alt. i): Set A and Set B are different (Set B is NOT a subset of Set A). Alt. ii): Set B is a subset of Set A (Set A and Set B are not the same). Alt. iii): Set A and Set B are the same.

- AI/ML model input consider: measurement results of K (K≥1) latest measurement instances with the following alternatives: Alt. 1): Only L1-RSRP measurement based on Set B; Alt 2): L1-RSRP measurement based on Set B and assistance information; Alt. 3): L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID.

- F predictions for F future time instances can be obtained based on the output of AI/ML model, where each prediction is for each time instance. At least F=1.

Set B is a set of beams whose measurements are taken as inputs of the AI/ML model.

Note: Beams in Set A and Set B can be in the same Frequency Range.

For both sub-use cases, the following alternatives are studied for the predicted beams:

- Alt.1: DL Tx beam prediction

- Alt.2: DL Rx beam prediction (deprioritized)

- Alt.3: Beam pair prediction (a beam pair consists of a DL Tx beam and a corresponding DL Rx beam)

Note: DL Rx beam prediction may or may not have spec impact.

The following alternatives according to AI/ML model output are considered:

- Alt.1: Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams

- e.g., N predicted beams can be the Top-N predicted beams

- Alt.2: Tx and/or Rx Beam ID(s) of the N predicted DL Tx and/or Rx beams and other information

- e.g., N predicted beams can be the Top-N predicted beams

- Alt.3: Tx and/or Rx Beam angle(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams

- e.g., N predicted beams can be the Top-N predicted beams

Notes: It is up to companies to provide other alternative(s). Beam ID is only used for discussion purposes. All the outputs are "nominal" and only for discussion purpose. The value of N is up to each company. All of the outputs in the above alternatives may vary based on whether the AI/ML model inference is at UE side or gNB side. The Top-N beam IDs might have been derived via post-processing of the ML-model output.

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, the necessity and potential BM-specific conditions/additional conditions for functionality(ies) and/or model(s) are considered at least from the following aspects:

- information regarding model inference

- Set A / Set B configuration

- performance monitoring

- data collection

- assistance information

For beam management use cases:

- For *model training*, training data can be generated by UE/gNB.

- For NW-side *model inference*, input data can be generated by UE and terminated at gNB.

- For UE-side *model inference*, input data is internally available at UE.

- For *performance* *monitoring* at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB.

## 5.3 Positioning accuracy enhancements

***Finalization of representative sub-use cases*:**

The following are selected as representative sub-use cases:

- Direct AI/ML positioning:

- AI/ML model output: UE location

- e.g., fingerprinting based on channel observation as the input of AI/ML model

- AI/ML assisted positioning:

- AI/ML model output: new measurement and/or enhancement of existing measurement

- e.g., LOS/NLOS identification, timing and/or angle of measurement, likelihood of measurement

More specifically, the following Cases are considered for the study:

- Case 1: UE-based positioning with UE-side model, direct AI/ML or AI/ML assisted positioning

- Case 2a: UE-assisted/LMF-based positioning with UE-side model, AI/ML assisted positioning

- Case 2b: UE-assisted/LMF-based positioning with LMF-side model, direct AI/ML positioning

- Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning

- Case 3b: NG-RAN node assisted positioning with LMF-side model, direct AI/ML positioning

One-sided model whose inference is performed entirely at the UE or at the network is prioritized in Rel-18 SI.

For all five positioning cases (Case 1/2a/2b/3a/3b), RAN1 has not considered prioritization.

For positioning enhancement use case:

- For *model training*, training data can be generated by UE/PRU/gNB/LMF.

- For LMF-side *model inference* (Case 2b, Case 3b), input data can be generated by UE/gNB and terminated at LMF.

- For gNB-side *model inference* (Case 3a), input data is internally available at gNB.

- For UE-side *model inference* (Case 1, Case 2a), input data is internally available at UE.

- For *performance monitoring* at the LMF side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE/gNB and terminated at LMF.

- For *performance monitoring* at the gNB side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by at least gNB.

# 6 Evaluations

In this clause, performance benefits of AI/ML based algorithms for the agreed use cases in the final representative set are evaluated.

The evaluation methodology is based on statistical models (from TR 38.901 and TR 38.857 for positioning), for link and system level simulations.

- Extensions of 3GPP evaluation methodology for better suitability to AI/ML based techniques should be considered as needed.

- Whether field data are optionally needed to further assess the performance and robustness in real-world environments should be discussed as part of the study.

- Need for common assumptions in dataset construction for training, validation and test for the selected use cases.

- Consider adequate model training strategy, collaboration levels and associated implications

- Consider agreed-upon base AI model(s) for calibration

- AI model description and training methodology used for evaluation should be reported for information and cross-checking purposes

Common KPIs and corresponding requirements for the AI/ML operations are to be determined. Also, use-case specific KPIs and benchmarks of the selected use-cases are to be determined.

- Performance, inference latency and computational complexity of AI/ML based algorithms should be compared to that of a state-of-the-art baseline

- Overhead, power consumption (including computational), memory storage, and hardware requirements (including for given processing delays) associated with enabling respective AI/ML scheme, as well as generalization capability should be considered.

## 6.1 Common evaluation methodology and KPIs

3GPP channel models (TR 38.901) are used as the baseline for evaluations. Note: additional results based on dataset other than that generated by 3GPP channel models are allowed.

**Common KPIs** (if applicable):

- Performance

- Intermediate KPIs

- Link and system level performance

- Generalization performance

- Over-the-air Overhead

- Overhead of assistance information

- Overhead of data collection

- Overhead of model delivery/transfer

- Overhead of other AI/ML-related signalling

- Inference complexity, including complexity for pre- and post-processing

- Computational complexity of model inference: TOPs, FLOPs, MACs

- there may be a disconnect between the actual complexity and the complexity evaluated as captured in clause 6 using these KPIs due to the platform-dependency and implementation (hardware and software) optimization solutions

- Computational complexity for pre- and post-processing

- Model complexity: e.g., the number of parameters and/or size (e.g., Mbyte)

- Complexity shall be reported in terms of "*number of real-value model parameters*" and "*number of real-value operations*" regardless of underlying model arithmetic

- Training complexity

- LCM related complexity and storage overhead

- Storage/computation for training data collection

- Storage/computation for training and model update

- Storage/computation for model monitoring

- Storage/computation for other LCM procedures, e.g., model activation, deactivation, selection, switching, fallback operation

For evaluation of performance monitoring approaches, the following model monitoring KPIs are considered as general guidance:

- Accuracy and relevance (i.e., how well does the given monitoring metric/methods reflect the model and system performance)

- Overhead (e.g., signalling overhead associated with model monitoring)

- Complexity (e.g., computation and memory cost for model monitoring)

- Latency (i.e., timeliness of monitoring result, from model failure to action, given the purpose of model monitoring)

Note: Other KPIs are not precluded. Relevant KPIs may vary across different model monitoring approaches.

## 6.2 CSI feedback enhancement

### 6.2.1 Evaluation assumptions, methodology and KPIs

For the performance evaluation of the AI/ML based CSI feedback enhancement, *system level simulation* approach is adopted as baseline. *Link level simulations* are optionally adopted.

For calibration purposes on the dataset and/or AI/ML model across companies, companies were encouraged to align the parameters (e.g., for scenarios/channels) for generating the dataset in the simulation as a starting point.

Performing intermediate evaluations on AI/ML model performance can be considered to derive the intermediate KPI(s) (e.g., accuracy of AI/ML output CSI) for the purpose of AI/ML solution comparison. If realistic DL channel estimation is considered, CSI accuracy is calculated using the target CSI from ideal channel and the output CSI from the realistic channel estimation. The target CSI from ideal channel equally applies to AI/ML based CSI feedback enhancement, and the baseline codebook.

***KPIs and Evaluation metrics*:**

- Capability/complexity: Floating point operations (FLOPs), AI/ML memory storage in terms of AI/ML model size and number of AI/ML parameters reported by companies who may select either or both

- Reported separately for the CSI generation part and the CSI reconstruction part (for CSI compression sub-use case)

- When reporting the computational complexity including the pre-processing and post-processing, the complexity metric of FLOPs may be reported separately for the AI/ML model and the pre/post processing. While reporting the FLOPs of pre-processing and post-processing the following boundaries are considered:

- Estimated raw channel matrix per each frequency unit as an input for pre-processing of the CSI generation part.

- Precoding vectors per each frequency unit as an output of post-processing of the CSI reconstruction part.

- CSI compression: Intermediate KPIs: SGCS and/or NMSE to evaluate the accuracy of the AI/ML output CSI

- For rank>1 cases, SGCS calculation/extension methods are to be reported:

- SGCS separately calculated for each layer (e.g., for K layers, K SGCS values are derived respectively, and comparison is performed per layer). Companies to ensure the correct calculation of SGCS and to avoid disorder issue of the output eigenvectors. Note: Eventual KPI can still be used to compare the performance.

- The granularity of the frequency unit for averaging operation is assumed to be:

- For 15kHz SCS: For 10MHz bandwidth: 4 RBs; for 20MHz bandwidth: 8 RBs

- For 30kHz SCS: For 10MHz bandwidth: 2 RBs; for 20MHz bandwidth: 4 RBs

- Other frequency unit granularities not precluded.

- CSI compression: Intermediate KPI: model monitoring mechanism is considered as:

- Step 1: Generate test dataset including K test samples.

- Step 2: For each of the K test samples, a bias factor of monitored intermediate KPI (KPI*Diff*) is calculated as a function of KPI*Diff* = *f* ( KPI*Actual* , KPI*Genie* ), where KPI*Actual* is the actual intermediate KPI, and KPI*Genie* is the genie-aided intermediate KPI.

- KPI*Diff* is considered for:

- Case 1: NW side monitoring of intermediate KPI, where the monitoring accuracy is evaluated for a given ground-truth CSI format (e.g., quantized ground-truth CSI with 8 bits scalar, R16 eType II-like method, etc.) or SRS measurements, where

- KPI*Actual* is calculated with the output CSI at the NW side and the given ground-truth CSI format or SRS measurements.

- KPI*Genie* is calculated with output CSI (as for KPI*Actual*) and the ground-truth CSI of Float32

- Note: if Float32 is used for KPI*Actual*, the monitoring accuracy is 100% if KPI*Actual* and KPI*Genie* are based on the same CSI sample.

- Case 2: UE side monitoring of intermediate KPI with a proxy model, where the monitoring accuracy is evaluated for the output of the proxy model at UE:

- Case 2-1: the proxy model is a proxy CSI reconstruction part, and KPI*Actual* is calculated based on the inference output of the proxy CSI reconstruction part at UE and the ground-truth CSI. Note: if the proxy CSI reconstruction model is the same as the actual CSI reconstruction model at the NW, the monitoring accuracy is 100%.

- Case 2-2: the proxy model directly outputs intermediate KPI (KPI*Actual*)

- KPI*Genie* is calculated with the output CSI at the NW side and the same ground-truth CSI.

- KPI*Diff* = *f* ( KPI*Actual* , KPI*Genie* ) can take the following forms:

- Option 1 (baseline for calibration): Gap between KPI*Actual* and KPI*Genie*, i.e. KPI*Diff* = (KPI*Actual* - KPI*Genie*); Monitoring accuracy is the percentage of samples for which | KPI*Diff*| < KPI*th 1*, where KPI*th 1* is a threshold of the intermediate KPI gap which can take the following values: 0.02, 0.05 and 0.1.

- Option 2 (optional and up to companies to report): Binary state where KPI*Actual* and KPI*Genie*, have different relationships to their threshold(s), i.e., KPI*Diff* = (KPI*Actual* > KPI*th 2*, KPI*Genie* < KPI*th 3*) OR (KPI*Actual* < KPI*th 2*, KPI*Genie* > KPI*th 3*), where KPI*th 2* is considered to be the same as KPI*th 3*. Monitoring accuracy is the percentage of samples for which KPI*Diff* = 0.

- Step 3: Calculate the statistical result of the KPI*Diff* over K test samples which represents the monitoring accuracy performance.

- Note: $KPI\_{Genie}$ is introduced for the evaluation and comparison purpose; it may not be available in the real network.

- Note: the complexity, overhead and latency of the monitoring scheme are to be reported.

- CSI prediction: Intermediate KPIs: calculated for each predicted instance if AI/ML model outputs multiple predicted instances

- If collaboration level x is reported as the benchmark, the EVM to distinguish level x and level y/z based AI/ML CSI prediction is considered from the generalization aspect, e.g., collaboration level y/z based CSI prediction is modelled as the fine-tuning case or generalization Case 1, while collaboration level x based CSI prediction is modelled as generalization Case 2 or Case 3.

- Throughput including: average UPT, 5%-ile UE throughput, and CDF of UPT

***Model generalization*:**

The following cases are considered for verifying the generalization performance of an AI/ML model over *various scenarios/configurations*:

- Case 1: The AI/ML model is trained based on training dataset from one Scenario#B/Configuration#B, and then the AI/ML model performs inference/test on a dataset from the same Scenario#B/Configuration#B

- Case 2: The AI/ML model is trained based on training dataset from a different data set than Scenario#B/Configuration#B, e.g., Scenario#A/Configuration#A, Scenario#B/Configuration#A, Scenario#A/Configuration#B, and then the AI/ML model performs inference/test on Scenario#B/Configuration#B.

- Case 3: The AI/ML model is trained based on training dataset constructed by mixing datasets from multiple scenarios/configurations including Scenario#B/Configuration#B and a different dataset than Scenario#B/Configuration#B, e.g., Scenario#A/Configuration#A, Scenario#B/Configuration#A, Scenario#A/Configuration#B, and then the AI/ML model performs inference/test on a dataset from a single Scenario/Configuration from Scenario#B/Configuration#B.

- Note: Companies to report the ratio for dataset mixing

- Note: number of the multiple scenarios/configurations can be larger than two

To verify the generalization performance of an AI/ML model over various scenarios, the *set of scenarios* are considered focusing on one or more of the following aspects:

- Various deployment scenarios (e.g., UMa, UMi, InH)

- Various outdoor/indoor UE distributions for UMa/UMi (e.g., 10:0, 8:2, 5:5, 2:8, 0:10)

- Various carrier frequencies (e.g., 2GHz, 3.5GHz)

- Other aspects of scenarios are not precluded, e.g., various antenna spacing, various antenna virtualization (TxRU mapping), various ISDs, various UE speeds, etc.

- Companies to report the selected scenarios for generalization verification

To verify the generalization/scalability performance of an AI/ML model over various configurations (e.g., which may potentially lead to different dimensions of model input/output), the *set of configurations* are considered focusing on one or more of the following aspects:

- Various bandwidths (e.g., 10MHz, 20MHz) and/or frequency granularities, (e.g., size of subband)

- Various sizes of CSI feedback payloads

- Various antenna port layouts, e.g., (N1/N2/P) and/or antenna port numbers (e.g., 32 ports, 16 ports)

- Various UE speeds (e.g., 10km/h, 30km/h, 60km/h, 120km/h, etc.) for CSI prediction sub use case

- Other aspects of configurations are not precluded, e.g., various numerologies, various rank numbers/layers, etc.

- The selected configurations for generalization verification

- The method to achieve generalization over various configurations to achieve scalability of the AI/ML input/output, including pre-processing, post-processing, etc

For evaluating the generalization/scalability over various configurations for **CSI compression**, to achieve the scalability over *different input/output dimensions*, companies to report which case(s) are evaluated from the following list:

- Case 0 (benchmark for comparison): One CSI generation part with fixed input and output dimensions to 1 CSI reconstruction part with fixed input and output dimensions for each of the different input and/or output dimensions.

- Case 1: One CSI generation part with scalable input and/or output dimensions to N>1 separate CSI reconstruction parts each with fixed and different output and/or input dimensions

- Case 2: M>1 separate CSI generation parts each with fixed and different input and/or output dimensions to one CSI reconstruction part with scalable output and/or input dimensions

- Case 3: A pair of CSI generation part with scalable input/output dimensions and CSI reconstruction part with scalable output and/or input dimensions

For CSI compression, to achieve the scalability over *different input dimensions* of CSI generation part (e.g., different bandwidths/frequency granularities, or different antenna ports), the generalization cases are elaborated as follows:

- Case 1: The AI/ML model is trained based on training dataset from a fixed dimension X1 (e.g., a fixed bandwidth/frequency granularity, and/or number of antenna ports), and then the AI/ML model performs inference/test on a dataset from the same dimension X1.

- Case 2: The AI/ML model is trained based on training dataset from a single dimension X1, and then the AI/ML model performs inference/test on a dataset from a different dimension X2.

- Case 3: The AI/ML model is trained based on training dataset by mixing datasets subject to multiple dimensions of X1, X2,..., Xn, and then the AI/ML model performs inference/test on a single dataset subject to the dimension of X1, or X2,…, or Xn.

- Note: For Case 2/3, the solutions to achieve the scalability between Xi and Xj, are reported by companies, including, e.g., pre-processing to angle-delay domain, padding, additional adaptation layer in AI/ML model, etc.

For CSI compression, to achieve the scalability over *different output dimensions* of CSI generation part (e.g., different generated CSI feedback dimensions), the generalization cases of are elaborated as follows

- Case 1: The AI/ML model is trained based on training dataset from a fixed output dimension Y1 (e.g., a fixed CSI feedback dimension), and then the AI/ML model performs inference/test on a dataset from the same output dimension Y1.

- Case 2: The AI/ML model is trained based on training dataset from a single output dimension Y1, and then the AI/ML model performs inference/test on a dataset from a different output dimension Y2.

- Case 3: The AI/ML model is trained based on training dataset by mixing datasets subject to multiple dimensions of Y1, Y2,..., Yn, and then the AI/ML model performs inference/test on a single dataset of Y1, or Y2,…, or Yn.

- Notes: For Case 1/2/3, companies to report whether the output of the CSI generation part is before quantization or after quantization. For Case 2/3, the solutions to achieve the scalability between Yi and Yj, are reported by companies, including, e.g., truncation, additional adaptation layer in AI/ML model, etc.

***Model Fine-tuning*:**

For the evaluation of the potential performance benefits of model fine-tuning of CSI feedback enhancement, which is optionally assessed, the following case is considered:

- The AI/ML model is trained based on training dataset from a different dataset than Scenario#B/Configuration#B, e.g., Scenario#A/Configuration#A, Scenario#A/Configuration#B, Scenario#B/Configuration#A, and then the AI/ML model is updated based on a fine-tuning dataset Scenario#B/Configuration#B. After that, the AI/ML model is tested on Scenario#B/Configuration#B.

- In this case, the fine-tuning dataset setting (e.g., size of dataset) is to be reported along with the improvement of performance.

***Further details on evaluations including training collaboration types***

For the evaluation of Type 2 (Joint training of the two-sided model at network side and UE side, respectively), following procedure is considered as an example:

- For each FP/BP loop,

- Step 1: UE side generates the FP results (i.e., CSI feedback) based on the data sample(s), and sends the FP results to NW side

- Step 2: NW side reconstructs the CSI based on FP results, trains the CSI reconstruction part, and generates the BP information (e.g., gradients), which are then sent to UE side

- Step 3: UE side trains the CSI generation part based on the BP information from NW side

- Note: the dataset between UE side and NW side is aligned.

- Other Type 2 training approaches are not precluded and reported by companies

For the evaluations of Type 2 (Joint training of the two-sided model at network side and UE side, respectively), the following evaluation cases are considered for *multi-vendors*,

- Case 1 (baseline): Type 2 training between one NW part model to one UE part model

- Case 2: Type 2 training between one NW part model and M>1 separate UE part models.

- Companies to report the AI/ML structures for the UE part model and the NW part model

- Case 3: Type 2 training between one UE part model and N>1 separate NW part models.

- Companies to report the AI/ML structures for the UE part model and the NW part model

For the evaluation of an example of Type 3 (Separate training at NW side and UE side), the following procedure is considered for the *sequential training starting with NW side training* (NW-first training):

- Step1: NW side trains the NW side CSI generation part (which is not used for inference) and the NW side CSI reconstruction part jointly

- Step2: After NW side training is finished, NW side shares UE side with a set of information (e.g., dataset) that is used by the UE side to be able to train the UE side CSI generation part

- Companies to report Dataset construction, e.g., the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only, or other information if applicable. Also report the Quantization behaviour, e.g., whether the shared output of the Network side CSI generation part is before or after quantization.

- Step3: UE side trains the UE side CSI generation part based on the received set of information

- Other Type 3 NW-first training approaches are not precluded

For the evaluation of an example of Type 3 (Separate training at NW side and UE side), the following procedure is considered for the *sequential training starting with UE side training* (UE-first training):

- Step1: UE side trains the UE side CSI generation part and the UE side CSI reconstruction part (which is not used for inference) jointly

- Step2: After UE side training is finished, UE side shares NW side with a set of information (e.g., dataset) that is used by the NW side to be able to train the CSI reconstruction part

- Companies to report Dataset construction, e.g., the set of information includes the input and label of the UE side CSI reconstruction part, or includes the input of the UE side CSI reconstruction part only, or other information if applicable. Also, report the Quantization behaviour, e.g., whether the shared input of the UE side CSI reconstruction part is before or after quantization.

- Step3: NW side trains the NW side CSI reconstruction part based on the received set of information

- Other Type 3 UE-first training approaches are not precluded

For the evaluation of an example of Type 3 (Separate training at NW side and UE side), the following evaluation cases for *sequential training* are considered for *multi-vendors*:

- Case 1 (baseline): Type 3 training between one NW part model and one UE part model

- Note 1: Case 1 can be naturally applied to the NW-first training case where 1 NW part model to M>1 separate UE part models

- Companies to report the dataset used between the NW part model and the UE part model, e.g., whether dataset for training UE part model is the same or a subset of the dataset for training NW part model

- Note 2: Case 1 can be naturally applied to the UE-first training case where 1 UE part model to N>1 separate NW part models

- Companies to report the dataset used between the NW part model and the UE part model, e.g., whether dataset for training NW part model is the same or a subset of the dataset for training UE part model

- Companies to report the AI/ML structures for the combination(s) of UE part model and NW part model, which can be the same or different

- Case 2: For UE-first training, Type 3 training between one NW part model and M>1 separate UE part models

- Note: Case 2 can be also applied to the M>1 UE part models to N>1 NW part models

- Companies to report the AI/ML structures for the M>1 UE part models and the NW part model

- Companies to report the dataset used at UE part models, e.g., same or different dataset(s) among M UE part models

- Case 3: For NW-first training, Type 3 training between one UE part model and N>1 separate NW part models

- Note: Case 3 can be also applied to the N>1 NW part models to M>1 UE part models

- Companies to report the AI/ML structures for the UE part model and the N>1 NW part models

- Companies to report the dataset used at NW part models, e.g., same or different dataset(s) among N NW part models

- Case 4: 1-on-1 training with joint training: benchmark/upper bound for performance comparison.

For the evaluation of Type 3 (Separate training at NW side and UE side), the following cases are considered for evaluations:

- Case 1 (baseline): Aligned AI/ML model structure between NW side and UE side

- Case 2: Not aligned AI/ML model structures between NW side and UE side

- Companies to report the AI/ML structures for the UE part model and the NW part model, e.g., different backbone (e.g., CNN, Transformer, etc.), or same backbone but different structure (e.g., number of layers)

- For the evaluation of training Type 3 under CSI compression, for the benchmark case (1-on-1 joint training) for performance comparison, the structures for the pair of NW part model/UE part model for the new case are the same with the Type 3 case to be compared, e.g., if the Type 3 is Transformer#1 for NW part model and CNN#1 for UE part model, then the benchmark case for performance comparison is also Transformer#1 for NW part model and CNN#1 for UE part model with joint training.

***Evaluation assumptions*:**

Table 6.2.1-1 presents the baseline system level simulation assumptions for AI/ML based CSI feedback enhancement evaluations.

Table 6.2.1-1: Baseline System Level Simulation assumptions for AI/ML based CSI feedback enhancement evaluations

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform | FDD (TDD is not precluded), OFDM |
| Multiple access | OFDMA |
| Scenario | Dense Urban (Macro only) is a baseline.Other scenarios (e.g., UMi@4GHz 2GHz, Urban Macro) are not precluded. |
| Frequency Range | FR1 only, 2GHz as baseline, optional for 4GHz (if R16 as baseline)FR1 only, 2GHz with duplexing gap of 200MHz between DL and UL, optional for 4GHz (if R17 as baseline) |
| Inter-BS distance | 200m |
| Channel model         | According to TR 38.901 |
| Antenna setup and port layouts at gNB | Companies need to report which option(s) are used between- 32 ports: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ- 16 ports: (8,4,2,1,1,2,4), (dH,dV) = (0.5, 0.8)λOther configurations are not precluded. |
| Antenna setup and port layouts at UE | 4RX: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ for (rank 1-4)2RX: (1,1,2,1,1,1,1), (dH,dV) = (0.5, 0.5)λ for (rank 1,2)Other configuration is not precluded. |
| BS Tx power | 41 dBm for 10MHz, 44dBm for 20MHz, 47dBm for 40MHz |
| BS antenna height | 25m |
| UE antenna height & gain | Follow TR36.873 |
| UE receiver noise figure | 9dB |
| Modulation | Up to 256QAM |
| Coding on PDSCH | LDPCMax code-block size=8448bit |
| Numerology | Slot/non-slot | 14 OFDM symbol slot |
| SCS | 15kHz for 2GHz, 30kHz for 4GHz |
| Simulation bandwidth | 10 MHz for 15kHz as a baseline, and configurations which emulate larger BW, e.g., same sub-band size as 40/100 MHz with 30kHz, may be optionally considered. Above 15kHz is replaced with 30kHz SCS for 4GHz (if R16 as baseline)20 MHz for 15kHz as a baseline (optional for 10 MHz with 15KHz), and configurations which emulate larger BW, e.g., same sub-band size as 40/100 MHz with 30kHz, may be optionally considered. Above 15kHz is replaced with 30kHz SCS for 4GHz (if R17 as baseline) |
| Frame structure | Slot Format 0 (all downlink) for all slots |
| MIMO scheme | SU/MU-MIMO with rank adaptation. Companies are encouraged to report the SU/MU-MIMO with RU.  |
| MIMO layers | For all evaluation, companies to provide the assumption on the maximum MU layers (e.g., 8 or 12) |
| CSI feedback | Feedback assumption at least for baseline scheme- CSI feedback periodicity (full CSI feedback): 5 ms (baseline)- Scheduling delay (from CSI feedback to time to apply in scheduling): 4 ms |
| Overhead | Companies shall provide the downlink overhead assumption (i.e., whether the CSI-RS transmission is UE-specific or not and take that into account for overhead computation) |
| Traffic model | At least, FTP model 1 with packet size 0.5 Mbytes is assumed.Other options are not precluded |
| Traffic load (Resource utilization) | 20/50/70%. Companies are encouraged to report the MU-MIMO utilization.  |
| UE distribution | CSI compression: 80% indoor (3 km/h), 20% outdoor (30 km/h)CSI prediction: 100% outdoor (10, 20, 30, 60, 120 km/h) including outdoor-to-indoor car penetration loss per TR 38.901 if the simulation assumes UEs inside vehicles. No explicit trajectory modeling considered for evaluations.  |
| UE receiver | MMSE-IRC as the baseline receiver |
| Feedback assumption | Realistic |
| Channel estimation          | Realistic as a baseline. Up to companies to choose the error modelling method for realistic channel estimation.Ideal DL channel estimation is optionally taken into the baseline of evaluation methodology for the purpose of calibration and/or comparing intermediate results (e.g., accuracy of AI/ML output CSI, etc.). Up to companies to report whether/how ideal channel is used in the dataset construction and performance evaluation/inference.Note: Eventual performance comparison with the benchmark release and drawing SI conclusions should be based on realistic DL channel estimation. |
| Evaluation Metric | Throughput and CSI feedback overhead as baseline metrics.The CSI feedback overhead is calculated as the weighted average of CSI payload per rank and the distribution of ranks reported by the UE. - For AI/ML based solutions: The above-mentioned "CSI feedback overhead" is calculated as max allowed bits at the given rank. - For legacy Type II CB: Option 2b is mandatorily reported by companies, while Option 2a can be optionally reported up to companies if partial NZC report is assumed for the legacy Type II CB- Option 2a: The above-mentioned "CSI feedback overhead" is calculated as each CSI reported payload with a given rank- Option 2b: The above-mentioned "CSI feedback overhead" is calculated as max allowed bits at the given rankAdditional metrics, e.g., ratio between throughput and CSI feedback overhead, can be used.Maximum overhead (payload size for CSI feedback)for each rank at one feedback instance is the baseline metric for CSI feedback overhead, and companies can provide other metrics. |
| Baseline for performance evaluation | For CSI compression:Companies need to report which option is used between:- Rel-16 TypeII Codebook as the baseline for performance and overhead evaluation.- Rel-17 TypeII Codebook as the baseline for performance and overhead evaluation.Additional assumptions from R17 TypeII EVM: Same consideration with respect to utilizing angle-delay reciprocity should be considered for the AI/ML based CSI feedback and the baseline scheme if R17 TypeII codebook is selected as baseline.Optionally, Type I Codebook (if it outperforms Type II Codebook) can be considered for comparing AI/ML schemes.For CSI-prediction: Both of the following are taken as baseline: - The nearest historical CSI without prediction- Non-AI/ML or AI/ML with collaboration Level x based CSI prediction for which corresponding details would need to be reportedNote: the specific non-AI/ML based CSI prediction is compatible with R18 MIMO; collaboration level x AI/ML based CSI prediction could be implementation based AI/ML compatible with R18 MIMO as an example.For the evaluation of CSI enhancements, companies can optionally provide the additional throughput baseline based on CSI without compression (e.g., eigenvector from measured channel), which is taken as an upper bound for performance comparison. |
| Note: the baseline EVM is used to compare the performance with the benchmark release, while the AI/ML related parameters (e.g., dataset construction, generalization verification, and AI/ML related metrics) can be of additional/different assumptions. The conclusions for the use cases in the SI should be drawn based on generalization verification over potentially multiple scenarios/configurations. |

Table 6.2.1-2 presents the baseline link level simulation assumptions for AI/ML based CSI feedback enhancement evaluations.

Table 6.2.1-2: Baseline Link Level Simulation assumptions for AI/ML based CSI feedback enhancement evaluations

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform  | FDD (TDD is not precluded), OFDM  |
| Carrier frequency | 2GHz as baseline, optional for 4GHz |
| Bandwidth | 10MHz or 20MHz |
| Subcarrier spacing | 15kHz for 2GHz, 30kHz for 4GHz |
| Nt | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ |
| Nr | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ |
| Channel model | CDL-C as baseline, CDL-A as optional |
| UE speed | 3kmhr, 10km/h, 20km/h or 30km/h to be reported by companies |
| Delay spread | 30ns or 300ns |
| Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline.Ideal DL channel estimation is optionally taken into the baseline of evaluation methodology for the purpose of calibration and/or comparing intermediate results (e.g., accuracy of AI/ML output CSI, etc.). Up to companies to report whether/how ideal channel is used in the dataset construction and performance evaluation/inference.Note: Eventual performance comparison with the benchmark release and drawing SI conclusions should be based on realistic DL channel estimation  |
| Rank per UE | Rank 1-4. Companies are encouraged to report the Rank number, and whether/how rank adaptation is applied |
| Note: the baseline EVM is used to compare the performance with the benchmark release, while the AI/ML related parameters (e.g., dataset construction, generalization verification, and AI/ML related metrics) can be of additional/different assumptions. The conclusions for the use cases in the SI should be drawn based on generalization verification over potentially multiple scenarios/configurations. |

***CSI compression sub use case specific aspects*:**

For the evaluation of the AI/ML based CSI compression sub use cases, a two-sided model is considered as a starting point, including an AI/ML-based CSI generation part to generate the CSI feedback information and an AI/ML-based CSI reconstruction part which is used to reconstruct the CSI from the received CSI feedback information. At least for inference, the CSI generation part is located at the UE side, and the CSI reconstruction part is located at the gNB side.

Figure 6.2.1-1 provides an example for the inference procedure for CSI compression. For generating the input of CSI generation model, it may need some further pre-processing on the measured channel; for the output of the CSI reconstruction model, some further post-processing may also be applied. Besides CSI feedback of quantization output, there may also be other CSI/PMI related information transmitted. There may be other examples of merging quantization/dequantization into the inference for CSI generation/reconstruction, CSI generation model/CSI reconstruction model, respectively.



Figure 6.2.1-1: An example of the CSI compression inference procedure

For the evaluation of the AI/ML based **CSI compression** sub use case, the following details of models are reported:

- The structure of the AI/ML model, e.g., type (CNN, RNN, Transformer, Inception, …), the number of layers, branches, real valued or complex valued parameters, etc.

- AI/ML model input (for CSI generation part)/output (for CSI reconstruction part) types for evaluations

- Data pre-processing/post-processing

- Loss function

- Specific quantization/dequantization method, e.g., vector quantization, scalar quantization, etc,

For the evaluation of the AI/ML based CSI compression sub use cases, at least the following types of AI/ML model input (for CSI generation part)/output (for CSI reconstruction part) are considered for evaluations:

- Raw channel matrix, e.g., channel matrix with the dimensions of Tx, Rx, and frequency unit. Companies to report the raw channel is in frequency domain or delay domain.

- Precoding matrix. Companies to report the precoding matrix is a group of eigenvector(s) or an eType II-like reporting (i.e., eigenvectors with angular-delay domain representation).

For the evaluation of quantization aware/non-aware training, the following cases are considered and reported by companies:

- Case 1: Quantization non-aware training, where the float-format variables are directly passed from CSI generation part to CSI reconstruction part during the training

o Fixed/pre-configured quantization method/parameters is applied for the inference phase. Companies to report the design of the fixed/pre-configured quantization method/parameters, e.g., quantization resolution, vector quantization codebook, etc

- Case 2: Quantization-aware training, where quantization/dequantization is involved in the training process

o Case 2-1: Fixed/pre-configured quantization method/parameters are applied during the training phase; the same quantization codebook is applied for the inference phase. Companies to report the design of the fixed/pre-configured quantization method/parameters, e.g., quantization resolution, vector quantization codebook, etc.

o Case 2-2: The quantization method/parameters are updated in together with the AI/ML models during the training; when training is finished, the final quantization codebook is applied for the inference phase. Companies to report how to update the quantization method/parameters during the training

- Quantization methods including uniform vs non-uniform quantization, scalar versus vector quantization, and associated parameters, e.g., quantization resolution, etc.

- How to use the quantization methods are reported by companies

Considering performance impact of ground-truth quantization in the CSI compression, study high resolution quantization methods for ground-truth CSI, including at least the following options:

- High resolution scalar quantization

- High resolution codebook quantization, e.g., Rel-16 TypeII-like method with new parameters, in which case companies are to report the R16 Type II parameters with specified or new/larger values to achieve higher resolution of the ground-truth CSI labels, e.g., L,$ p\_{v}$, $β$, reference amplitude, differential amplitude, phase, etc

- Float32 adopted as the baseline/upper-bound for performance comparisons

- Consider legacy values of PC6 & PC8 for performance comparison

For CSI compression sub use case with rank ≥ 1, AI/ML model setting to adapt to ranks/layers to be reported amongst the following options:

- Option 1-1 (rank specific): Separated AI/ML models are trained per rank value and applied for corresponding ranks to perform individual inference, any specific model operates on multi-layers jointly.

- Option 1-2 (rank common): A unified AI/ML model is trained and applied for adaptive ranks to perform inference, the model operates on multi-layers jointly.

- Option 2 (layer specific): Separated AI/ML models are trained per layer value and applied for corresponding layers to perform individual inference.

o Note: input/output type is Precoding matrix

o Companies to report the setting is

 Option 2-1: layer specific and rank common (different models applied for different layers; for a specific layer, the same model is applied for all rank values), or

 Option 2-2: layer specific and rank specific (different models applied for different layers; for a specific layer, different models are applied for different rank values)

- Option 3 (layer common): A unified AI/ML model is trained and applied for each layer to perform individual inference.

o Note: input/output type is Precoding matrix

o Companies to report whether the setting is

 Option 3-1: layer common and rank common (A unified AI/ML model is applied for each layer under any rank value to perform individual inference), or

 Option 3-2: layer common and rank specific (different models applied for different rank values; for a specific rank, the same model is applied for all layers)

For CSI compression sub use case with rank >1, for a given configured Max rank=K, the complexity of FLOPs is reported as the maximum FLOPs over all ranks each includes the summation of FLOPs for inference per layer if applicable, e.g.,

- Option 1-1 (rank specific): Max FLOPs over K rank specific models.

- Option 1-2 (rank common): FLOPs of the rank common model.

- Option 2-1 (layer specific and rank common): Sum of the FLOPs of K models (for the rank=K).

- Option 2-2 (layer specific and rank specific): Max of the FLOPs over K ranks, k=1,…K, each with a sum of k models.

- Option 3-1 (layer common and rank common): K \* FLOPs of the common model.

- Option 3-2 (layer common and rank specific): Max of the FLOPs over K ranks, k=1,…K, each with k \* FLOPs of the layer common model.

For CSI compression sub use case with rank >1, the storage of memory storage/number of parameters is reported as the summation of memory storage/number of parameters over all models potentially used for any layer/rank, e.g.,

- Option 1-1 (rank specific)/Option 3-2 (layer common and rank specific): Sum of memory storage/number of parameters over all rank specific models.

- Option 1-2 (rank common): A single memory storage/number of parameters for the rank common model.

- Option 2-1 (layer specific and rank common): Sum of memory storage/number of parameters over all layer specific models.

- Option 2-2 (layer specific and rank specific): Sum of memory storage/number of parameters for the specific models over all ranks and all layers in per rank.

- Option 3-1 (layer common and rank common): A single memory storage/number of parameters for the common model

For the evaluation of CSI compression, the specific CQI determination method(s) for AI/ML can be reported by introducing an additional field in the template, e.g.,

- Option 1a: CQI is calculated based on the target CSI from the realistic channel estimation.

- Option 1b: CQI is calculated based on the target CSI from the realistic channel estimation and potential adjustment.

- Option 1c: CQI is calculated based on traditional codebook.

- Option 2a: CQI is calculated based on CSI reconstruction output, if CSI reconstruction model is available at the UE and UE can perform reconstruction model inference with potential adjustment.

o Option 2a-1: The CSI reconstruction part for CQI calculation at the UE same as the actual CSI reconstruction part at the NW.

o Option 2a-2: The CSI reconstruction part for CQI calculation at the UE is a proxy model, which is different from the actual CSI reconstruction part at the NW.

- Option 2b: CQI is calculated using two stage approach, UE derives CQI using precoded CSI-RS transmitted with a reconstructed precoder.

***CSI prediction sub use case specific aspects*:**

Figure 6.2.1-2 provides an example for the inference procedure for CSI prediction. For generating the input of CSI prediction model, it may need some further pre-processing on the measured channel; for the output of the CSI prediction model, some further post-processing may also be applied.



Figure 6.2.1-2: An example of the CSI prediction inference procedure

For the evaluation of the AI/ML based **CSI prediction** sub use case, the following details of models are reported:

- The structure of the AI/ML model, e.g., type (FCN, RNN, CNN,…), the number of layers, branches, format of parameters, etc.

- The input CSI type, e.g., raw channel matrix, eigenvector(s) of the raw channel matrix, feedback CSI information, etc.

- Including assumptions on the observation window, i.e., number/time distance of historic CSI/channel measurements

- The output CSI type, e.g., channel matrix, eigenvector(s), feedback CSI information, etc.

- Including assumptions on the prediction window, i.e., number/time distance of predicted CSI/channel

- Data pre-processing/post-processing

- Loss function

For the input CSI type, both of the following types are considered for evaluations:

- Raw channel matrices

- Eigenvector(s)

For SLS, spatial consistency Procedure A with 50m decorrelation distance from TR 38.901 is used (if not used, assumptions used need to be reported). UE velocity vector is assumed as fixed over time in Procedure A modelling.

### 6.2.2 Performance results

CSI\_Table 1 through CSI\_Table 7 in attached Spreadsheets for CSI feedback enhancement evaluations present the performance results for:

- CSI\_Table 1. Evaluation results for CSI compression of 1-on-1 joint training without model generalization/scalability

- CSI\_Table 2. Evaluation results for CSI compression with model generalization

- CSI\_Table 3. Evaluation results for CSI compression with model scalability

- CSI\_Table 4. Evaluation results for CSI compression of multi-vendor joint training without model generalization/scalability

- CSI\_Table 5. Evaluation results for CSI compression of separate training without model generalization/scalability

- CSI\_Table 6. Evaluation results for CSI prediction without model generalization/scalability

- CSI\_Table 7. Evaluation results for CSI prediction with model generalization

For the evaluation of CSI compression of 1-on-1 joint training without model generalization/scalability, the following baselines are recommended to facilitate calibration of results:

- Benchmark: R16 eType II CB;

- Others can be additionally submitted, e.g., Type I CB.

- Input/Output type: Eigenvectors of the current CSI

- Other can be additionally submitted, e.g., eigenvectors with additional past CSI, eType II-like input, raw channel matrix, etc.

- Ground-truth CSI quantization method: Float32, i.e., without quantization (baseline/upper-bound for performance comparison)

- Other high resolution CSI quantization methods can be additionally submitted for comparison, e.g., R16 eType II-like method with new parameters, scalar quantization, etc.

- Rank/layer adaptation settings for rank>1: Option 3-1, i.e., layer common and rank common.

- Other rank>1 options can be additionally submitted for comparison, e.g., Option 1-1/1-2/2-1/2-2/3-2.

- Quantization method: quantization-aware training (Case 2-1 or Case 2-2)

- Quantization non-aware training can be additionally submitted for comparison

- SQ and/or VQ is up to companies; companies are encouraged to provide results of various cases for comparison.

- Performance metric for intermediate KPI: SGCS

- NMSE can be additionally submitted

The CSI feedback reduction is provided for three CSI feedback overhead ranges (CSI feedback overhead A, CSI feedback overhead B, CSI feedback overhead C), where for each CSI feedback overhead range of the benchmark, it is calculated as the gap between the CSI feedback overhead of benchmark and the CSI feedback overhead of AI/ML corresponding to the same mean UPT. The various CSI feedback overhead ranges are defined as:

 CSI feedback overhead A such that A ≤ β ˖80 bits

 CSI feedback overhead B such that β ˖100 bits ≤ B ≤ β ˖140 bits

 CSI feedback overhead C ≥ β ˖230 bits

 where, β = 1 for rank = 1 and β = 1.5 for rank > 1

Note: companies report the exact CSI feedback overhead considered.

Note: the CSI feedback overhead reduction and gain for mean/5%tile UPT are determined at the same payload size for benchmark scheme.

Note: "Benchmark" means the type of Legacy CB used for comparison. "Quantization/dequantization method" includes the description of training awareness (Case 1/2-1/2-2), type of quantization/dequantization (SQ/VQ), etc. "Input type" means the input of the CSI generation part. "Output type" means the output of the CSI reconstruction part.

For the evaluation of *CSI prediction* without model generalization/scalability verification, the following baselines are recommended to facilitate calibration of results:

- UE speed: 10km/h, 30km/h, 60km/h;

- Others can be additionally submitted, e.g., 120km/h.

- Input/Output type: Raw channel matrix

- Other can be additionally submitted, e.g., eigenvectors.

- Observation window (number/distance): 5/5ms, 10/5ms

- Other observation window configurations can be additionally submitted for comparison, e.g., 3/5ms, 4/5ms, 8/2.5ms, 10/4ms, etc.

- Prediction window (number/distance between prediction instances/distance from the last observation instance to the 1st prediction instance): 1/5ms/5ms

- Other prediction window configurations can be additionally submitted for comparison, e.g., 3/5ms/5ms, 5/5ms/5ms, 4/2.5ms/2.5ms, 5/4ms/4ms, etc.

- Performance metric for intermediate KPI: SGCS

- NMSE can be additionally submitted.

- Spatial consistency configuration (optional): procedure A with 50m decorrelation distance and channel updating periodicity of 1 ms.

For the evaluation of *CSI prediction* with model generalization/scalability verification, the following baselines are recommended to facilitate calibration of results:

- Performance metric for intermediate KPI: SGCS

- NMSE can be additionally submitted.

#### 6.2.2.1 1-on-1 joint training for CSI compression

***Input/output type***

For the evaluation of CSI compression, for the type of AI/ML model input (for CSI generation part)/output (for CSI reconstruction part), a vast majority of companies adopt precoding matrix as model input/output.

Note: For the evaluations of CSI compression with 1-on-1 joint training, 22 sources take precoding matrix without angular-delay domain conversion as the model input/output; 2 sources take precoding matrix with angular-delay domain representation as the model input/output. No company submitted explicit channel matrix as input.

The complexity metric in terms of FLOPs and number of parameters of AI/ML models adopted in the evaluations of CSI compression are summarized in Figure 6.2.2.1-1, where the complexity for the CSI generation part and the complexity for the CSI reconstruction part are illustrated separately.

- A majority of 25 sources adopt the CSI generation model subject to the FLOPs from 10M to 800M, and 26 sources adopt the CSI reconstruction model subject to the FLOPs from 10M to 1100M.

- A majority of 21 sources adopt the CSI generation model subject to the number of parameters from 1M to 13M, and 22 sources adopt the CSI reconstruction model subject to the number of parameters from 1M to 17M.

- Results refer to Table 1 of clause 7.3, R1-2310450.



Figure 6.2.2.1-1: Complexity of AI/ML models from evaluation results in terms of FLOPs and number of parameters for CSI compression

***SGCS performance***

For the evaluation of AI/ML based CSI compression compared to the *benchmark in terms of SGCS*,

For Max rank 1, Layer 1,

- 14 sources observe the performance gain of 2.6%~ 8.8% at CSI payload X (small payload);

- 18 sources observe the performance gain of 0.9%~ 8.1% at CSI payload Y (medium payload);

- 16 sources observe the performance gain of 0.9%~ 7% at CSI payload Z (large payload);

- Note: 3 sources observe the performance gain of 0%, 10.2%~11.6% at CSI payload X (small payload), 0.9% at CSI payload Y (medium payload), -0.3% at CSI payload Z (large payload) which biases from the majority range.

For Max rank 2, Layer 1,

- 15 sources observe the performance gain of 3.9%~ 11% at CSI payload X (small payload);

- 13 sources observe the performance gain of 0.7%~ 4.5% at CSI payload Y (medium payload);

- 14 sources observe the performance gain of -0.2%~ 6.5% at CSI payload Z (large payload);

- Note: 4 sources observe the performance gain of 12.7%~15.6% at CSI payload X (small payload), 5%~10.6% at CSI payload Y (medium payload), 7.1% at CSI payload Z (large payload) which biases from the majority range.

For Max rank 2, Layer 2, more gains are observed in general compared with Layer 1 of Max rank 2:

- 13 sources observe the performance gain of 5.92%~ 30.2% at CSI payload X (small payload);

- 13 sources observe the performance gain of 1.5%~ 23.08% at CSI payload Y (medium payload);

- 11 sources observe the performance gain of 4.4%~ 12.99% at CSI payload Z (large payload);

- Note: 5 sources observe the performance gain of -7.4%~1.1%, 49.3% at CSI payload X (small payload), -0.3%~1.5%, 41.7% at CSI payload Y (medium payload), -0.4%~2.2%, 45.9% at CSI payload Z (large payload) which biases from the majority range.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2.

- CSI payload X is ≤ 80 bits; CSI payload Y is 100 bits - 140 bits; CSI payload Z is ≥ 230 bits; where X, Y, Z are applicable per layer.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.6 of R1-2308340.

***Mean UPT for FTP traffic***

For the evaluation of AI/ML based CSI compression compared to the *benchmark in terms of mean UPT* *under FTP* traffic, more gains are achieved by Max rank 2 compared with Max rank 1 in general:

- For Max rank 1, in general the performance gain increases with the increase of RU:

- For RU≤39%, 7 sources observe the performance gain of 0.2%~2%

- 6 sources observe the performance gain of 0.29%~2% at CSI feedback overhead A (small overhead);

- 6 sources observe the performance gain of 0.2%~1% at CSI feedback overhead B (medium overhead)

- 4 sources observe the performance gain of 0.33%~1% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 7 sources observe the performance gain of 0.1%~4%

- 5 sources observe the performance gain of 1.09%~3% at CSI feedback overhead A (small overhead);

- 4 sources observe the performance gain of 0.80%~2% at CSI feedback overhead B (medium overhead);

- 7 sources observe the performance gain of 0.1%~4% at CSI feedback overhead C (large overhead);

- For RU≥70%, 9 sources observe the performance gain of 0.23%~9%

- 9 sources observe the performance gain of 0.38%~9% at CSI feedback overhead A (small overhead)

- 8 sources observe the performance gain of 0.62%~5% at CSI feedback overhead B (medium overhead)

- 8 sources observe the performance gain of 0.23%~6% at CSI feedback overhead C (large overhead);

- Note: 5 sources observe gain of 0.1%~0.2%, 1.7%~2.51% at RU≤39%, 0.5%~1%, 2.34%~21.21% at RU 40%-69%, 2.51%~21.5% at RU≥70%, which bias from the majority ranges.

- For Max rank 2, in general the performance gain increases with the increase of RU:

- For RU≤39%, 8 sources observe the performance gain of -0.3%~6%

- 7 sources observe the performance gain of 1%~6% at CSI feedback overhead A (small overhead);

- 7 sources observe the performance gain of 0.5%~6% at CSI feedback overhead B (medium overhead);

- 8 sources observe the performance gain of -0.3%~6% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 10 sources observe the performance gain of -0.5%~10%

- 8 sources observe the performance gain of 3%~10% at CSI feedback overhead A (small overhead);

- 8 sources observe the performance gain of 1.2%~9% at CSI feedback overhead B (medium overhead)

- 10 sources observe the performance gain of -0.5%~9% at CSI feedback overhead C (large overhead)

- For RU≥70%, 11 sources observe the performance gain of -0.2%~15%

- 11 sources observe the performance gain of 5%~15% at CSI feedback overhead A (small overhead);

- 11 sources observe the performance gain of 3%~9% at CSI feedback overhead B (medium overhead);

- 10 sources observe the performance gain of -0.2%~12% at CSI feedback overhead C (large overhead);

- Note: 5 sources observe gain of 0.3%, 7%~30% at RU≤39%, 1%, 18%~23% at RU 40%-69%, 12.71%~26.8% at RU≥70%, which bias from the majority ranges.

- For Max rank 4:

- For RU≤39%, 2 sources observe the performance gain of -4%~6%

- 2 sources observe the performance gain of 2.5%~6% at CSI feedback overhead A (small overhead);

- 1 source observes the performance gain of 6% at CSI feedback overhead B (medium overhead);

- 2 sources observe the performance gain of -4%~0% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 3 sources observe the performance gain of -1.8%~12.22%

- 3 sources observe the performance gain of 3%~12.22% at CSI feedback overhead A (small overhead);

- 2 sources observe the performance gain of 7.04%~11% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of -1.8%~8.19% at CSI feedback overhead C (large overhead);

- For RU≥70%, 3 sources observe the performance gain of -1%~17%

- 3 sources observe the performance gain of 3%~17% at CSI feedback overhead A (small overhead);

- 2 sources observe the performance gain of 6.64%~17% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of -1%~8.40% at CSI feedback overhead C (large overhead);

- Note: 1 source observes significant gain or significant loss under Max rank 4 due to specific CQI/RI selection method (e.g., Option 1a/2a) for AI/ML and/or CQI/RI determination method for eType II benchmark.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is mean UPT for Max rank 1, Max rank 2, or Max rank 4.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.12 of R1-2308342.

***5% UPT for FTP traffic***

For the evaluation of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under FTP*, more gains are achieved by Max rank 2 compared with Max rank 1 in general:

- For Max rank 1, in general the performance gain increases with the increase of RU:

- For RU≤39%, 3 sources observe the performance gain of 0.8%~3%

- 3 sources observe the performance gain of 1.72%~3% at CSI feedback overhead A (small overhead);

- 3 sources observe the performance gain of 0.80%~1.2% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of 1.68%~3% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 6 sources observe the performance gain of 0.1%~7%

- 6 sources observe the performance gain of 2.8%~7% at CSI feedback overhead A (small overhead);

- 3 sources observe the performance gain of 1.22%~2.7% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of 0.1%~3.25% at CSI feedback overhead C (large overhead);

- For RU≥70%, 8 sources observe the performance gain of 0.85%~20.43%

- 8 sources observe the performance gain of 4%~20.43% at CSI feedback overhead A (small overhead);

- 7 sources observe the performance gain of 1%~10.13% at CSI feedback overhead B (medium overhead);

- 8 sources observe the performance gain of 0.85%~8% at CSI feedback overhead C (large overhead);

- Note: 4 sources observe gain of 0% and 5.6%~5.7% at RU≤39%, 4.2%~5.8% at RU 40%-69%, 23%~50% at RU≥70%, which bias from the majority ranges.

- For Max rank 2, in general the performance gain increases with the increase of RU:

- For RU≤39%, 8 sources observe the performance gain of -2%~5%

- 5 sources observe the performance gain of 1.1%~5% at CSI feedback overhead A (small overhead);

- 6 sources observe the performance gain of -2%~3% at CSI feedback overhead B (medium overhead);

- 7 sources observe the performance gain of -0.5%~5% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 8 sources observe the performance gain of -4%~13%

- 6 sources observe the performance gain of 7%~13% at CSI feedback overhead A (small overhead);

- 7 sources observe the performance gain of 0.3%~8% at CSI feedback overhead B (medium overhead);

- 6 sources observe the performance gain of -4%~8% at CSI feedback overhead C (large overhead);

- For RU≥70%, 9 sources observe the performance gain of -1.3%~24%

- 6 sources observe the performance gain of 10.26%~24% at CSI feedback overhead A (small overhead);

- 6 sources observe the performance gain of 9%~15.02% at CSI feedback overhead B (medium overhead);

- 8 sources observe the performance gain of -1.3%~13.67% at CSI feedback overhead C (large overhead);

- Note: 7 sources observe gain of 4.4%~13% at RU≤39%, -8%~-2%, 10%~25.6% at RU 40%-69%, -10%~-8.1% at RU≥70%, which bias from the majority ranges.

- For Max rank 4:

- For RU≤39%, 2 sources observe the performance gain of -1.6%~10%

- 2 sources observe the performance gain of 8%~10% at CSI feedback overhead A (small overhead);

- 1 source observes the performance gain of 5% at CSI feedback overhead B (medium overhead);

- 2 sources observe the performance gain of -1.6%~1% at CSI feedback overhead C (large overhead);

- For RU 40%-69%, 3 sources observe the performance gain of -1.7%~23%

- 3 sources observe the performance gain of 5%~17% at CSI feedback overhead A (small overhead);

- 2 sources observe the performance gain of 6.17%~23% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of -1.7%~9.47% at CSI feedback overhead C (large overhead);

- For RU≥70%, 3 sources observe the performance gain of 2%~31%

- 3 sources observe the performance gain of 5.8%~31% at CSI feedback overhead A (small overhead);

- 2 sources observe the performance gain of 10.2%~30% at CSI feedback overhead B (medium overhead);

- 3 sources observe the performance gain of 2%~15% at CSI feedback overhead C (large overhead);

- Note: 1 source observes significant gain or significant loss under Max rank 4 due to specific CQI/RI selection method (e.g., Option 1a/2a) for AI/ML and/or CQI/RI determination method for eType II benchmark

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is 5% UPT for Max rank 1, Max rank 2, or Max rank 4.

- Benchmark is Rel-16 Type II codebook.

- Results refer to Table 5.13 of R1-2308342.

***Mean UPT for full buffer***

For the evaluation of AI/ML based CSI compression compared to the *benchmark, in terms of mean UPT under full buffer*, more gains are achieved by Max rank 2 compared with Max rank 1 in general:

- For Max rank 1, 8 sources observe the performance gain of 1.1%~11%

- 6 sources observe the performance gain of 6%~11% at CSI feedback overhead A (small overhead);

- 6 sources observe the performance gain of 3%~7% at CSI feedback overhead B (medium overhead);

- 8 sources observe the performance gain of 1.1%~11% at CSI feedback overhead C (large overhead);

- For Max rank 2, 9 sources observe the performance gain of 0.2%~15%

- 9 sources observe the performance gain of 4%~15% at CSI feedback overhead A (small overhead);

- 9 sources observe the performance gain of 2%~10% at CSI feedback overhead B (medium overhead);

- 9 sources observe the performance gain of -0.2%~14% at CSI feedback overhead C (large overhead);

- Note: For Max rank 4, 1 source observes gain of 7.44%~9.95% over CSI feedback overhead A/B/C.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.7 of R1-2308340.

***5% UPT for full buffer***

For the evaluation of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under full buffer*,

- For Max rank 1, 5 sources observe the performance gain of 0%~20.9%

- 5 sources observe the performance gain of 2.5%~20.9% at CSI feedback overhead A (small overhead);

- 5 sources observe the performance gain of 2.3%~17.4% at CSI feedback overhead B (medium overhead);

- 4 sources observe the performance gain of 0%~6.62% at CSI feedback overhead C (large overhead);

- For Max rank 2, 6 sources observe the performance gain of -7%~14.9%

- 6 sources observe the performance gain of 4.1%~14.9% at CSI feedback overhead A (small overhead);

- 5 sources observe the performance gain of 0.3%~4% at CSI feedback overhead B (medium overhead);

- 6 sources observe the performance gain of -7%~6.03% at CSI feedback overhead C (large overhead);

- Note: For Max rank 4, 1 source observes gain of 3.59%~6.15% over CSI feedback overhead A/B/C.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.8 of R1-2308340.

***CSI feedback reduction***

For the evaluation of AI/ML based CSI compression, compared to the benchmark, in terms of CSI feedback reduction,

- For Max rank = 1,

- For CSI feedback overhead A (small overhead), 1 source observes the CSI feedback reduction of 10.24% for FTP traffic;

- For CSI feedback overhead B (medium overhead), 3 sources observe the CSI feedback reduction of 15.62%~60% for FTP traffic, and 2 sources observe the CSI feedback reduction of 37%~66% for full buffer;

- For CSI feedback overhead C (large overhead), 2 sources observe the CSI feedback reduction of 14.37%~55% for FTP traffic, and 2 sources observes the CSI feedback reduction of 50%~53% for full buffer;

- Note: For CSI feedback overhead C (large overhead), 1 source observes CSI feedback reduction of 75% for FTP traffic.

- For Max rank = 2,

- For CSI feedback overhead A (small overhead), 3 sources observe the CSI feedback reduction of 20.83%~54% for FTP traffic, and 1 source observes the CSI feedback reduction of 56% for full buffer;

- For CSI feedback overhead B (medium overhead), 3 sources observe the CSI feedback reduction of 22.22%~52% for FTP traffic, and 2 sources observe the CSI feedback reduction of 52% for full buffer;

- For CSI feedback overhead C (large overhead), 3 sources observe the CSI feedback reduction of 10%~58.33% for FTP traffic, and 2 sources observe the CSI feedback reduction of 22%~54% for full buffer;

- Note: For CSI feedback overhead B (medium overhead), 1 source observes CSI feedback reduction of up to ~83% for FTP traffic using particular VQ codebook solution.

- For Max rank = 4,

- For CSI feedback overhead A (small overhead), 2 sources observe the CSI feedback reduction of 50%~79% for FTP traffic, and 1 source observes the CSI feedback reduction of 70.53% for full buffer;

- For CSI feedback overhead B (medium overhead), 2 sources observe the CSI feedback reduction of 36.10%~78% for FTP traffic, and 1 source observes the CSI feedback reduction of 47.74% for full buffer;

- For CSI feedback overhead C (large overhead), 2 sources observe the CSI feedback reduction of 8%~58% for FTP traffic, and 1 source observes the CSI feedback reduction of 42.59% for full buffer;

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is CSI feedback overhead reduction for Max rank 1/2/4.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.30 of R1-2308344.

***Monitoring for intermediate KPI, NW side monitoring***

For the evaluation of intermediate *KPI based monitoring* mechanism for CSI compression, for monitoring Case 1, in terms of monitoring accuracy with Option 1,

- For ground-truth CSI format of R16 eType II CB, monitoring accuracy is increased with the increase of the resolution for the ground-truth CSI (number of bits for each sample of ground-truth CSI) in general, with the impact of increased overhead, wherein

- for ground-truth CSI format of R16 eType II CB with PC#6, 4 sources observe KPIDiff as 13.2%~71.6%/ 28.5%~100%/ 68.4%~100% for KPIth\_1=0.02/0.05/0.1, respectively.

- Note: two sources observed averaging on the test samples improves the monitoring accuracy.

- for ground-truth CSI format of R16 eType II CB with PC#8, 5 sources observe KPIDiff as 21%~43.0%/ 48.1%~79.1%/ 79.8%~97.1% for KPIth\_1=0.02/0.05/0.1, respectively.

- for ground-truth CSI format of R16 eType II CB with new parameter of 580-750bits CSI payload size, 2 sources observe KPIDiff as 35.4%~63%/ 77.9%~93.0%/ 99.5%~99.9% for KPIth\_1=0.02/0.05/0.1, respectively, which have 12.7%~20%/ 13.9%~29.8%/ 8%~31.1% gain over PC#8.

- for ground-truth CSI format of R16 eType II CB with new parameter of around 1000bits CSI payload size, 4 sources observe KPIDiff as 34.9%~89%/ 82.9%~100%/ 99.9%~100% for KPIth\_1=0.02/0.05/0.1, respectively, which have 12.2%~68%/ 18%~43.62%/ 2.9%~31% gain over PC#8 from 3 sources and 4.67%~10.6%/ 0%~5.88%/ 0%~0.49% gain over PC#6 from 1 source.

- for ground-truth CSI format of R16 eType II CB with new parameter of around 1600bits CSI payload size, 2 sources observe KPIDiff as 89.1%~97%/ 99.9%~100%/ 100% for KPIth\_1=0.02/0.05/0.1, respectively, which have 76%/33%/3% gain over PC#8 from 1 source.

- For ground-truth CSI format of 4 bits scalar quantization, 2 sources observe KPIDiff as 9.4%~47%/ 96.3%~100%/ 100% for KPIth\_1=0.02/0.05/0.1, respectively.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Time independency is assumed over the test samples for monitoring

- Precoding matrix is used as the model input.

- 1-on-1 joint training is assumed.

- The performance metric is monitoring accuracy for Layer 1.

- Note: Results refer to Table 5.21 of R1-2308343.

***Monitoring for intermediate KPI, UE side monitoring***

For the evaluation of intermediate *KPI based monitoring* mechanism for CSI compression, for Case 2, in terms of monitoring accuracy with Option 1,

- For Case 2-1 subject to generalization Case 1 for the proxy model, 5 sources observe KPIDiff as 31%~84%/ 65.63%~99.8%/ 95%~100% for KPIth\_1=0.02/0.05/0.1, respectively;

- Compared with monitoring Case 1 with ground-truth CSI format of R16 eType II CB with new parameter of around 1000bits CSI payload size,

- 2 sources observe +0.99%~+4.07% gain at KPIth\_1=0.02;

- 3 sources observe -6.03%~-58%/ -0.2%~-24%/ 0%~-5% degradation for KPIth\_1=0.02/0.05/0.1, respectively;

- Compared with monitoring Case 1 with ground-truth CSI format of R16 eType II CB with new parameter of around 1600bits CSI payload size, 2 sources observe -16.35%~-66%/ -0.4%~-24%/ 0%~-24% degradation for KPIth\_1=0.02/0.05/0.1, respectively.

- Note: For Case 2-1 subject to generalization Case 2 for the proxy model, 2 sources observe -1.77%~-37.42% / -1.07%~-23.93%/ -0.16%~-14% compared with generalization Case 1 with the same testing scenario.

- Note: For Case 2-2, 1 source observes KPIDiff as 61%~72.1%/ 91.2%~96.6%/ 99.2%~99.75% under generalization Case 1 for the proxy model, and 60%~71.3%/ 90.4%~99.3%/ 99%~100% under generalization Case 3 for the proxy model, for KPIth\_1=0.02/0.05/0.1, respectively.

- Note: for Case 2-1, 1 source observes that if different model backbone is adopted for proxy model as compared to the NW part model, it has negative impact to the monitoring performance.

- Note: for the complexity and overhead analysis:

- Case 2-1/Case 2-2 have smaller air-interface overhead for UE report for monitoring compared with Case 1. Overhead of proxy model from LCM perspective, if any, is not evaluated.

- The complexity aspect for Case 1, Case 2-1 and Case 2-2 is not evaluated.

- Note: "Generalization Case 1" means the proxy model is trained based on training dataset from one Scenario#B, and then tested for monitoring on a dataset from the same Scenario#B. "Generalization Case 2" means the proxy model is trained based on training dataset from one Scenario#A, and then tested for monitoring on a dataset from a different Scenario#B. "Generalization Case 3" means the proxy model is trained based on mixing datasets from multiple scenarios including Scenario#B, and then tested for monitoring on the dataset from Scenario#B.

- Note: two sources observed averaging on the test samples improves the monitoring accuracy.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Time independency is assumed over the test samples for monitoring.

- Precoding matrix is used as the model input.

- 1-on-1 joint training is assumed.

- The performance metric is monitoring accuracy for Layer 1.

- Note: Results refer to Table 5.22 of R1-2308343.

***Quantization methods, quantization awareness for training***

For the comparison of *quantization methods* for CSI compression, *quantization non-aware training* (Case 1) is in general inferior to the *quantization aware training* (Case 2-1/2-2), and may lead to lower performance than the benchmark:

- For scalar quantization, compared with benchmark,

- -2.4%~-43.2% degradations are observed for quantization non-aware training (Case 1) from 6 sources.

- 3.9%~8.64% gains are observed for quantization aware training with fixed/pre-configured quantization method/parameters (Case 2-1) from 5 sources, which are 17.3%~83.2% gains over quantization non-aware training (Case 1) from 5 sources and 7.56%~11.55% gains over quantization non-aware training (Case 1) from 1 source.

- Note: 0.72% gains are observed for Case 2-1 from 1 source due to SQ parameter chosen without matching latent distribution, which achieves 13.9% gains over Case 1.

- 8.91% gains are observed for quantization aware training with jointly updated quantization method/parameters (Case 2-2) from 1 source, which are 23.1% gains over quantization non-aware training (Case 1) from 1 source.

- For vector quantization, compared with benchmark,

- -2%~-10% degradations are observed for quantization non-aware training (Case 1) from 1 source.

- 5.64%~7.55% gains are observed for quantization aware training with fixed/pre-configured quantization method/parameters (Case 2-1) from 3 sources, which are 3%~21.6% gains over quantization non-aware training (Case 1) from 3 sources.

- 4.6%~13.01% gains are observed for quantization aware training with jointly updated quantization method/parameters (Case 2-2) from 7 sources, which are 10.7%~30% gains over quantization non-aware training (Case 1) from 4 sources and 3.66%~9.8% gains over quantization non-aware training (Case 1) from 2 sources.

- In general, Case 2-2 outperforms Case 2-1 with 0.46%~5.1% gains, as observed by 6 sources.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS for Layer 1.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.14 of R1-2308342.

***Quantization methods, quantization format***

For the comparison of *quantization methods* for CSI compression, in general vector quantization (VQ) has comparable performance with scalar quantization (SQ):

- For SQ and VQ under the same training case, it is

- observed by 3 sources that VQ under Case 2-1 has -1%~-4.5% degradation over SQ under Case 2-1,

- observed by 1 source that VQ under Case 2-1 has 1.1% gain over SQ under Case 2-1, and

- observed by 3 sources that VQ under Case 2-2 has 0.7%~3.8% gain over SQ under Case 2-2.

- Note: VQ under Case 2-1 has 8% gains over SQ under Case 2-1 as observed from 1 source due to SQ parameter chosen without matching latent distribution.

- For SQ and VQ across training cases, it is

- observed by 6 sources that VQ under Case 2-2 has 0.46%~4% gain over SQ under Case 2-1, and

- observed by 1 source that VQ under Case 2-2 has -1.3% degradation over SQ under Case 2-1.

- observed by 1 source that VQ under Case 2-1 has -2.9%~-6.4% degradation over SQ under Case 2-2.

- Note: in general, more companies observing gain of VQ over SQ than companies observing loss.

- Note: it is observed by 1 source that combined SQ and VQ under Case 2-2 has minor gain of 0.2% over VQ only under Case 2-2.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS for Layer 1.

- Benchmark is Rel-16 Type II codebook.

- Note: Results refer to Table 5.15 of R1-2308342.

***High resolution ground-truth CSI for training***

For the evaluation of high-resolution quantization of the ground-truth CSI for the training of CSI compression, compared to the upper-bound of Float32, quantized high resolution ground-truth CSI can achieve significant overhead reduction with minor performance loss if the parameters are appropriately selected.

- For high resolution scalar quantization,

- Float16 achieves 50% overhead reduction and -0.6% or less performance loss from 2 sources

- 8 bits scalar quantization achieves 75% overhead reduction and -0.14%~-0.9% performance loss from 2 sources

- For high resolution R16 eType II-like quantization,

- R16 eType II CB with legacy parameters can achieve significant overhead reduction while with performance loss compared to Float32, wherein:

- PC#6 achieves around 99% overhead reduction with -1.4% ~-1.7% performance loss from 2 sources, and -3%~-9.5% performance loss from 4 sources.

- PC#8 achieves around 98% overhead reduction with 0% ~-1.7% performance loss from 3 sources, and -2.9%~-5.5% performance loss from 5 sources.

- For R16 eType II CB with new parameters:

- R16 eType II CB with new parameter of 1000-1400bits CSI payload size achieves 95%~97.5% overhead reduction (3~4.1 times overhead compared to PC8) with performance gain of 0.7%~4.3% over PC#8 from 4 sources.

- R16 eType II CB with new parameter of 1500-2100bits CSI payload size achieves 94%~96.2% overhead reduction (4.8~6.1 times overhead compared to PC8) with performance gain of 1.3%~5.4% over PC#8 from 3 sources.

- Note: it is observed by 1 source that using R16 eType II-like quantization with legacy PC may achieve close performance to Float32 by dataset dithering.

- Note: the new parameters include at least one from the follows:

- L= 8, 10, 12;

- pv = 0.8, 0.9, 0.95;

- reference amplitude = 6 bits, 8 bits; differential amplitude = 4bits; phase = 5 bits, 6 bits;

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS for Layer 1.

- Note: Results refer to Table 5.18 of R1-2308342.

#### 6.2.2.2 Generalization evaluations for CSI compression

***Generalization over deployment scenarios***

From the results for the *generalization verification* of AI/ML based CSI compression *over various deployment scenarios* compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain deployment scenario#B and applied for inference with a same deployment scenario#B,

- For *generalization Case 2*, generalized performance may be achieved for certain combinations of deployment scenario#A and deployment scenario#B but not for others:

- If deployment scenario#A is UMi & deployment scenario#B is UMa, deployment scenario#A is UMa & deployment scenario#B is UMi, or deployment scenario#A is UMa & deployment scenario#B is InH:

- 14 sources observe that generalized performance can be achieved:

- For deployment scenario#A is UMi & deployment scenario#B is UMa, 9 sources observe less than -1.6% degradation or positive gain.

- For deployment scenario#A is UMa & deployment scenario#B is UMi, 10 sources observe less than -1.5% degradation or positive gain.

- For deployment scenario#A is UMa & deployment scenario#B is InH, 2 sources observe less than -0.6% degradation or positive gain.

- 13 sources observe that moderate/significant degradations are suffered under generalization Case 2:

- For deployment scenario#A is UMi & deployment scenario#B is UMa, 10 sources observe -1.69%~-21.1% degradation.

- For deployment scenario#A is UMa & deployment scenario#B is UMi, 9 sources observe -1.7%~-8.1% degradation.

- For deployment scenario#A is UMa & deployment scenario#B is InH, 3 sources observe -1.74%~-31.6% degradation.

- If deployment scenario#A is InH & deployment scenario#B is Uma/UMi, significant performance degradations are observed under generalization Case 2:

- For deployment scenario#A is InH & deployment scenario#B is UMa, 5 sources observe -5.55%~ -27.7% degradation.

- For deployment scenario#A is InH & deployment scenario#B is UMi, 3 sources observe -8.63%~-20% degradation

- For *generalization Case 3*, generalized performance of the AI/ML model can be achieved (0%~-4% loss or positive gain) for deployment scenario#B subject to any of UMa, UMi, and InH, if the training dataset is constructed with data samples subject to multiple deployment scenarios including deployment scenario#B, as observed by 15 sources.

- Minor loss (0%~-1.6%) are observed by 15 sources.

- Moderate loss (-1.69%~-4%) are observed by 8 sources.

- Positive gains are observed by 10 sources.

- Note: Significant degradations of up to -6.7% are observed by 2 sources for deployment scenario#B subject to UMa, and by 2 sources for deployment scenario#B subject to UMi.

- Note: For generalization Case 2, if deployment scenario#A is UMi & deployment scenario#B is InH, 3 sources observe different trends, where significant performance degradations of -27.8%~-32.86% are observed by two sources, while moderate performance degradations of -1.44%~-2.41% are observed by another source.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1/2.

- Note: Results refer to Table 5.1 of R1-2308340.

***Generalization over UE distributions***

For the *generalization verification* of AI/ML based CSI compression *over various UE distributions* compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain UE distribution#B and applied for inference with a same UE distribution#B,

- For generalization Case 2, generalized performance may be achieved for some certain combinations of UE distribution#A and UE distribution#B but not for others

- If UE distribution#A is Outdoor & UE distribution#B is Indoor, 7 sources observe that moderate/significant degradations of -1.9%~-11.5% degradation are suffered,

- Note: 1 source observes minor degradation of -0.48%~-0.93% for partial cases.

- If UE distribution#A is Indoor & UE distribution#B is Outdoor, 7 sources observe minor loss of less than -1.11% degradation or positive gain

- For generalization Case 3, generalized performance of the AI/ML model can be achieved (0%~-1.54% loss or positive gain) for UE distribution#B subject to any of Outdoor and Indoor, if the training dataset is constructed with data samples subject to multiple UE distributions including UE distribution#B, as observed by 6 sources.

- Minor loss (0%~-1.54%) are observed by 5 sources.

- Positive gains are observed by 4 sources.

- Note: Moderate degradations of up to -3.9% are still observed by 2 sources for UE distribution#B subject to Indoor.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1/2.

- Note: Results refer to Table 5.9 of R1-2308340.

***Generalization over carrier frequencies***

For the *generalization verification* of AI/ML based CSI compression *over various carrier frequencies* compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain carrier frequency#B and applied for inference with a same carrier frequency#B,

- For generalization Case 2, generalized performance may be achieved in general

- If carrier frequency#A is 3.5/4GHz & carrier frequency#B is 2GHz, 3 sources observe generalized performance of less than -0.8% degradation.

- If carrier frequency#A is 2GHz & carrier frequency#B is 3.5/4GHz, 5 sources observe generalized performance of less than -1.06% degradation or positive gain.

- Note: 2 sources observes significant degradations up to -6.6%.

- For generalization Case 3, generalized performance of the AI/ML model may be achieved (0%~-1.2% loss or positive gain) for carrier frequency#B subject to any of 2GHz and 3.5/4GHz, if the training dataset is constructed with data samples subject to multiple carrier frequencies including carrier frequency#B, as observed by 4 sources.

- Minor loss (0%~-1.2%) are observed by 4 sources.

- Positive gains are observed by 4 sources.

- Note: Significant degradations of up to -4.9% are still observed by 1 source for carrier frequency#B subject to 3.5/4GHz

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1.

- Antenna layouts are assumed as the same over the different frequency carriers.

- Note: Results refer to Table 5.2 of R1-2308340.

***Generalization over TxRU mappings***

For the generalization verification of AI/ML based CSI compression over various TxRU mappings, compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain TxRU mapping#B and applied for inference with a same TxRU mapping#B,

- For generalization Case 2, significant degradations are suffered in general from the perspective of the layouts of antenna ports, as observed by 2 sources:

- For TxRU mapping#A is [2,8,2] & TxRU mapping#B is [4,4,2] or TxRU mapping#A is [8,2,2] & TxRU mapping#B is [4,4,2], 2 sources observe -13%~-36.1% degradation.

- For TxRU mapping#A is [4,4,2] & TxRU mapping#B is [2,8,2] or TxRU mapping#A is [8,2,2] & TxRU mapping#B is [2,8,2], 2 sources observe -7%~-23.6% degradation.

- For TxRU mapping#A is [4,4,2] & TxRU mapping#B is [8,2,2] or TxRU mapping#A is [2,8,2] & TxRU mapping#B is [8,2,2], 1 source observes -19%~-27% degradation.

- For generalization Case 2, generalized performance may be achieved for some certain combinations of TxRU mapping#A and TxRU mapping#B but not for others, from the perspective of the layouts of antenna element mapping, as observed by 2 sources:

- For TxRU mapping#A is 8x8x2 & TxRU mapping#B is 2x8x2, 2 sources observe minor/moderate degradation of -0.6%~-2.5%.

- For TxRU mapping#A is 2x8x2 & TxRU mapping#B is 8x8x2, 1 source observes moderate degradation of -3%.

- For generalization Case 3, generalized performance of the AI/ML model can be achieved (0%~-4.4% loss or positive gain) for TxRU mapping#B subject to any of [2,8,2], [4,4,2], and [8,2,2] from the perspective of the layouts of antenna ports, or subject to any of 8x8x2 and 2x8x2 from the perspective of the layouts of antenna element mapping, if the training dataset is constructed with data samples subject to TxRU mappings including TxRU mapping#B, as observed by 4 sources.

- Minor loss (0%~-2%) are observed by 4 sources.

- Moderate loss (-2.5%~-4.4%) are observed by 1 source.

- Positive gains are observed by 1 source.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1.

- [x,y,z] for TxRU mapping: Vertical port number, Horizontal port number, polarization

- AxBxC for TxRU mapping: AxBxC antenna elements virtualized to [2,8,2]

- Note: Results refer to Table 5.19 of R1-2308342.

#### 6.2.2.3 Scalability evaluations for CSI compression

***Scalability over CSI payload sizes***

For the scalability verification of AI/ML based CSI compression *over various CSI payload sizes*, compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain CSI payload size#B and applied for inference with a same CSI payload size#B,

- For generalization Case 2, significant performance degradations are observed in general, as -5.3%~-14.7% degradations are observed by 2 sources.

- Generalized performance of the AI/ML model can be achieved (-0%~-5.9%loss) under generalization Case 3 for the inference on CSI payload size#B, if the training dataset is constructed with data samples subject to multiple CSI payload sizes including CSI payload size#B, and an appropriate scalability solution is performed to scale the dimension of the AI/ML model, shown by 13 sources (10 sources showing -0%~-2.2% loss, 7 sources showing -2.3%~-5.9% loss, 5 sources showing positive gain). The scalability solution is adopted as follows:

- Pre/post-processing of truncation/padding, adopted by 6 sources, showing -0% ~-5.9% loss or positive gain.

- Various quantization granularities, adopted by 1 source, showing -0.7% loss or positive gain.

- Adaptation layer in the AL/ML model, adopted by 6 sources, showing -0%~-4.78% loss or positive gain.

- Note: Significant degradations of up to -14.22% are still observed by 2 sources for generalization Case 3.

- Generalized performance of the AI/ML model can also be achieved by finetuning models on CSI payload size#B, showing loss 0%~-2.2% by 2 sources

The above results are based on the following assumptions:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- Input/output scalability dimension Case 3 is adopted: A pair of CSI generation part with scalable input/output dimensions and CSI reconstruction part with scalable output and/or input dimensions.

- The performance metric is SGCS in linear value for layer 1/2.

- Note: Results refer to Table 5.10 of R1-2308340.

***Scalability over bandwidths***

For the scalability verification of AI/ML based CSI compression *over various bandwidths*, compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain bandwidth#B and applied for inference with a same bandwidth#B,

- For generalization Case 2, if bandwidth#A is 20MHz & bandwidth#B is 10MHz, or bandwidth#A is 10MHz & bandwidth#B is 20MHz, or bandwidth#A is 10MHz & bandwidth#B is 5MHz:

- 2 sources observe that generalized performance can be achieved:

- For bandwidth#A is 20MHz & bandwidth#B is 10MHz, 1 source observes less than -1.28% degradation.

- For bandwidth#A is 10MHz & bandwidth#B is 20MHz, 2 sources observe less than -1.1% degradation.

- 1 source observe that moderate/significant degradations are suffered under generalization Case 2:

- For bandwidth#A is 10MHz & bandwidth#B is 5MHz, 1 source observes larger than -2.5% degradation.

- For generalization Case 3, 3 sources observe that generalized performance of the AI/ML model can be achieved (0%~-2.97% loss) for bandwidth#B subject to each of 10MHz/52RB and 20MHz and 48RB, if the training dataset is constructed with data samples subject to multiple bandwidths including bandwidth#B.

- Minor loss (0%~-1.7%) are observed by 2 sources.

- Moderate loss (-1.91%~-2.97%) are observed by 2 sources.

- Positive gains are observed by 2 sources.

- Note: Significant loss (-5.4%) is observed by 1 source.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1/2.

- Note: Results refer to Table 5.31 of R1-2308344.

***Scalability over Tx port numbers***

For the *scalability verification* of AI/ML based CSI compression *over various Tx port numbers* compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain Tx port number#B and applied for inference with a same Tx port number#B,

- For generalization Case 2, significant performance degradations are observed in general, if Tx port number#A is 32 & Tx port number#B is 16, as -3.37%~-21.8% degradations are observed by 4 sources

- For generalization Case 3, generalized performance of the AI/ML model can be achieved (0%~-3.94% loss or positive gains) for Tx port number#B subject to any of 16 and 32, if the training dataset is constructed with data samples subject to multiple Tx port numbers including Tx port number#B, and an appropriate scalability solution is performed to scale the dimension of the AI/ML model, as observed by 9 sources.

- Minor loss (0%~-1.6%) are observed by 8 sources.

- Moderate loss (-2.02%~-3.94%) are observed by 4 sources.

- Positive gains are observed by 5 sources.

- Note: Significant degradations of up to -9.76% are still observed by 2 sources for deployment scenario#B subject to 32 ports, and for deployment scenario#B subject to 16 ports

- Note: Pre/post-processing of truncation/padding is adopted by 6 sources, and adaptation layer in the AL/ML model is adopted by 1 source.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- The performance metric is SGCS in linear value for layer 1/2/3/4.

- Note: Results refer to Table 5.3 of R1-2308340.

#### 6.2.2.4 Multi-vendor joint training for CSI compression

***1 NW part model to M>1 UE part models***

For the evaluation of Type 2 training between 1 NW part model and M>1 separate UE part models (Case 2), as compared to joint training between 1 NW part model and the 1 UE part model,

- 7 sources observe minor degradation of -0%~-1.67% or positive gain;

- 3 sources observe moderate degradation of -2.5%~-6.5%.

- Note: among the above sources, 5 sources adopt simultaneous training, while 1 source adopts sequential training starting with NW side training.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1.

- Same pair of NW part model and UE part model between 1-on-1 joint training and Type 2 training.

- M=2, 3, or 4 are considered.

- Note: Results refer to Table 5.23 of R1-2308343.

***1 UE part model to N>1 NW part models***

For the evaluation of Type 2 training between 1 UE part model and N>1 separate NW part models (Case 3), as compared to joint training between 1 NW part model and the 1 UE part model,

- 2 sources observe minor degradation of -0%~-0.8% or positive gain;

- 1 source observe moderate degradation of -1.4%~-4.2%.

- Note: among the above sources, 1 source adopts simultaneous training.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1.

- Same pair of NW part model and UE part model between 1-on-1 joint training and Type 2 training.

- N=2, 3, or 4 are considered.

- Note: Results refer to Table 5.24 of R1-2308343.

#### 6.2.2.5 Separate training for CSI compression

***NW first training, 1 NW part model to 1 UE part model, same backbone***

For the evaluation of Type 3 *NW first separate training with dataset sharing* manner for CSI compression for the pairing of 1 NW to 1 UE (Case 1), as compared to 1-on-1 joint training between the NW part model and the UE part model,

- For the NW first separate training case where the *same backbone* is adopted for both the NW part model and the UE part model, minor degradation is observed for both the cases where the shared output of the Network side CSI generation part is before or after quantization:

- For the case where the shared output of the Network side CSI generation part is after quantization, 9 sources observe -0%~-0.5% degradation, 10 sources observe -0.5%~-1% degradation, and 2 sources observe -1%~-1.3% degradation.

- For the case where the shared output of the Network side CSI generation part is before quantization, 6 sources observe -0%~-0.8% degradation, and 1 source observes -1%~-1.5% degradation.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement "the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1/2.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and NW first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- Note: Results refer to Table 5.16 of R1-2308342.

***Impact of shared dataset under 1 NW part model to 1 UE part model***

For the evaluation of Type 3 NW/UE first separate training with dataset sharing manner for CSI compression for the pairing of 1 NW to 1 UE (Case 1), as compared to the case where the same set of dataset is applied for training the NW part model and training the UE part model, if the dataset#2 applied for training the UE/NW part model is a subset of the dataset#1 applied for training the NW/UE part model,

- If the dataset#2 is appropriately selected, minor additional performance degradation can be achieved, as -0%~-0.59% gap is observed from 3 sources.

- If the dataset#2 has a significantly reduced size compared to dataset#1, moderate/significant additional performance degradation may occur, as -0.6%~-4.83% gap is observed from 4 sources.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement where "the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1/2.

- Note: Results refer to Table 5.4 of R1-2308340.

***NW first training, 1 NW part model to 1 UE part model, different backbones***

For the evaluation of Type 3 NW first separate training with dataset sharing manner for CSI compression, for the pairing of 1 NW to 1 UE (Case 1), as compared to 1-on-1 joint training between the NW part model and the UE part model,

- For the NW first separate training case where different backbones are adopted for the NW part model and the UE part model, more degradations are observed in general than the situation where the same backbone is adopted for the NW part model and the UE part model.

- For the case where the shared output of the Network side CSI generation part is after quantization, 3 sources observe minor degradation of -0%~-1.02%, and 3 sources observe moderate degradation of -1.46%~-5.1%.

- For the case where the shared output of the Network side CSI generation part is before quantization, 2 sources observe minor degradation of -0%~-0.1%, 1 source observes moderate degradation of -2.03%.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement where "the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1/2.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and NW first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- Note: Results refer to Table 5.16 of R1-2308342.

***NW first training, 1 UE part model to N>1 NW part models***

For the evaluation of Type 3 NW first separate training with dataset sharing manner for CSI compression, for the pairing between 1 UE part model and N>1 separate NW part models (Case 3), when taking 1-on-1 joint training between the NW part model and the UE part model as benchmark, larger performance loss is observed in general than the case of NW first separate training with 1 UE part model and 1 NW part model pairing (Case 1):

- 6 sources observe minor loss of -0%~-1.6% compared to the 1-on-1 joint training.

- 3 sources observe moderate loss of -1.9%~-6.64% compared to the 1-on-1 joint training.

- 5 sources observe significant loss of -37.9%~-87% compared to the 1-on-1 joint training.

- Note: as opposed to companies which observe significant loss, the minor loss observed by other companies may due to the fact that special handling (e.g., adaptation layer) is performed to pair with N>1 NW part models during the training at the UE side.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement, where "the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and NW first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- N=2, 3, or 4 are considered.

- Note: Results refer to Table 5.20 of R1-2308342.

***UE first training, 1 NW part model to 1 UE part model, same backbone***

For the evaluation of Type 3 UE first separate training with dataset sharing manner for CSI compression for the pairing of 1 NW to 1 UE (Case 1), as compared to 1-on-1 joint training between the NW part model and the UE part model,

- For the UE first separate training case where the same backbone is adopted for both the UE part model and the NW part model, minor degradation is observed in general for both the cases where the shared input of the UE side CSI reconstruction part is before or after quantization:

- For the case where the shared input of the UE side CSI reconstruction part is after quantization, 9 sources observe -0%~-0.42% degradation, 2 sources observe -0.7%~-0.9% degradation, and 3 sources observe -1.05%~-1.8% degradation.

- For the case where the shared input of the UE side CSI reconstruction part is before quantization, 3 sources observe -0%~-0.8% degradation, and 2 sources observe -1.3%~-2.9% degradation.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement where "the set of information includes the input and label of the UE side CSI reconstruction part, or includes the input of the UE side CSI reconstruction part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1/2.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and UE first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- Note: Results refer to Table 5.17 of R1-2308342.

***UE first training, 1 NW part model to 1 UE part model, different backbones***

For the evaluation of Type 3 UE first separate training with dataset sharing manner for CSI compression, for the pairing of 1 NW to 1 UE (Case 1), as compared to 1-on-1 joint training between the NW part model and the UE part model,

- For the UE first separate training case where different backbones are adopted for the NW part model and the UE part model, more degradations are observed in general than the situation where the same backbone is adopted for the NW part model and the UE part model.

- For the case where the shared input of the UE side CSI reconstruction part is after quantization, 5 sources observe minor degradation of -0.23%~-1.07%, and 1 source observes moderate degradation of -1.74%~-1.88%.

- For the case where the shared input of the UE side CSI reconstruction part is before quantization, 1 source observes moderate degradation of -1.58%~-2.73%.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement, where "the set of information includes the input and label of the UE side CSI reconstruction part, or includes the input of the UE side CSI reconstruction part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1/2.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and UE first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- Note: Results refer to Table 5.17 of R1-2308342.

***UE first training, M>1 UE part models to 1 NW part model***

For the evaluation of Type 3 UE first separate training with dataset sharing manner for CSI compression, for the pairing between M>1 separate UE part models and 1 NW part model (Case 2), when taking 1-on-1 joint training between the NW part model and the UE part model as benchmark, larger performance loss is observed in general than the case of UE first separate training with 1 UE part model and 1 NW part model pairing (Case 1):

- 8 sources observe minor loss of -0%~-1.82% compared to 1-on-1 joint training.

- 4 sources observe moderate loss of -2.17%~-4.96% compared to 1-on-1 joint training.

- 2 sources observe significant loss of -11.56%~-73.7% compared to 1-on-1 joint training.

- Note: 1 source observes other UE first separate training implementations may achieve better performance.

- Note: the dataset sharing behaviour from above sources follows the example of the agreement, where "the set of information includes the input and output of the Network side CSI generation part, or includes the output of the Network side CSI generation part only".

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS for Layer 1.

- Same size of training dataset for benchmark, NW part training and the UE part training

- Same pair of NW part model and UE part model between 1-on-1 joint training and UE first separate training.

- Quantization/dequantization method/parameters between NW side and UE side are aligned.

- M=2, 3, or 4 are considered.

- Note: Results refer to Table 5.25 of R1-2308343.

#### 6.2.2.6 Basic performance for CSI prediction

The complexity values in terms of FLOPs and number of parameters of AI/ML models adopted in the evaluations of CSI prediction are summarized in Figure 6.2.2.6-1.

- Results refer to Table 2 of clause 7.3, R1-2310450.



Figure 6.2.2.6-1: Complexity of AI/ML models from evaluation results in terms of FLOPs
and number of parameters for CSI prediction.

***SGCS performance, impact of input type***

For the AI/ML based CSI prediction, compared with the benchmark of the nearest historical CSI:

- spatial consistency is not adopted in 15 sources, wherein:

- 15 sources observe the gain of 0.46% ~ 44.8% using raw channel matrix as input, wherein

- 4 sources observe the gain of 0.46%~6.3%.

- 14 sources observe the gain of 7.57%~26.47%.

- 5 sources observe the gain of 29.03%~44.8%.

- 4 sources observe the gain of 2.24% ~ 19.4% using precoding matrix as input, which is in general worse than using raw channel matrix as input

- spatial consistency is adopted in 4 sources, all of which use raw channel matrix as input, wherein

- 3 sources observe the gain of 1.7%~35.51%.

- 1 source observe the gain of 76.6%.

- 1 source observe the loss of -5.5%.

The above results are based on the following assumptions:

- The observation window considers to start as early as 15ms~50ms.

- A future 4ms or 5ms instance from the prediction output is considered for calculating the metric.

- UE speed includes 10km/h, 30km/h, and 60km/h. The same fixed UE speed is assumed for both training and inference.

- The performance metric is SGCS in linear value for layer 1.

- Note: Results refer to Table 5.26 of R1-2308344.

***SGCS performance, impact of UE speed***

For the AI/ML based CSI prediction, compared to the Benchmark#1 of the nearest historical CSI, *in terms of SGCS*, from UE speed perspective, in general the gain of AI/ML based solution is related with the UE speed:

- For 10km/h UE speed, 6 sources observe 2.4%~12.5% gain (2.4%~12.5% gain for 5 sources who do not adopt spatial consistency, and 8.7% gain for 1 source who adopts spatial consistency), 1 source observes 21.93% gain (who does not adopt spatial consistency).

- For 30km/h UE speed, 1 source observes loss of -5.5% (who adopts spatial consistency), 3 sources observe 6%~10.43% gain (who do not adopt spatial consistency), 8 sources observe 12.65%~33% gain (14.65%~33% gain for 7 sources who do not adopt spatial consistency, and 12.65% gain for 1 source who adopts spatial consistency), and 3 sources observe 41.75%~ 76.6% gain (41.75%~ 44.8% gain for 2 sources who do not adopt spatial consistency, and 76.6% gain for 1 source who adopts spatial consistency), which are in general larger than 10km/h UE speed.

- For 60km/h UE speed, 3 sources observe 0.46%~2.6% gain (0.46%~2.3% gain for 2 sources who do not adopt spatial consistency, and 1.7%~2.6% gain for 1 source who adopts spatial consistency), 7 sources observe 9.1%~20.6% gain (9.1%~20.6% gain for 6 sources who do not adopt spatial consistency, and 13.8% gain for 1 source who adopts spatial consistency), 1 source observe 29.03% gain, which are in general smaller than 30km/h UE speed.

The above results are based on the following assumptions:

- The observation window considers to start as early as 15ms~50ms.

- A future 4ms or 5ms instance from the prediction output is considered for calculating the metric.

- Raw channel matrix is considered as model input

- The performance metric is SGCS in linear value for layer 1.

- No post processing is considered.

- The same fixed UE speed is assumed for both training and inference.

- Note: Results refer to Table 5.27 of R1-2308344.

***SGCS performance, impact of observation window***

For the AI/ML based CSI prediction, compared to the Benchmark#1 of the nearest historical CSI, *in terms of SGCS*, from observation window length perspective, in general the gain of AI/ML based solution is slightly increased with the increase of the length for the observation window:

- When the observation window is increased from 5/5ms to 8/5ms, the gain over benchmark is increased by 0.28%~2.19%, as observed by 2 sources.

- When the observation window is increased from 5/5ms to 15/5ms, the gain over benchmark is increased by 5.59%~10.32%, as observed by 1 source.

- When the observation window is increased from 4/5ms to 8/5ms and 10/5ms, the gain over benchmark is increased by 0.96%~4.23% and 1%~4.42%, respectively, as observed by 2 sources.

The above results are based on the following assumptions:

- The UE speed is 30km/h.

- A future 4ms or 5ms instance from the prediction output is considered for calculating the metric.

- Raw channel matrix is considered as model input

- The performance metric is SGCS in linear value for layer 1.

- No post processing is considered.

- Note: Results refer to Table 5.32 of R1-2308344.

***SGCS performance, impact of prediction window***

- For the AI/ML based CSI prediction, compared to the Benchmark#1 of the nearest historical CSI, *in terms of SGCS*, from prediction window length perspective, in general the gain of AI/ML based solution is related with the prediction length in terms of the distance to the applicable time of the predicted CSI:

- When the prediction length is increased from 10ms to 15ms, the gain over benchmark is reduced (gap from -1.13%~-51%), as observed by 3 sources.

- When the prediction length is increased from 2.5ms/3ms to 5ms, the gain over benchmark is increased (gap from +5.85%~+13%), as observed by 2 sources.

- When the prediction length is increased from 5ms to 10ms, 5 sources observe the gain over benchmark is reduced (gap from -1%~-12.1%) while 2 sources observe the gain over benchmark is increased (+11.65%~+45.5%).

The above results are based on the following assumptions:

- The UE speed is 30km/h.

- The observation window considers to start as early as 15ms~50ms.

- Raw channel matrix is considered as model input.

- The performance metric is SGCS in linear value for layer 1.

- No post processing is considered.

- Note: Results refer to Table 5.33 of R1-2308344.

***Mean UPT***

For the AI/ML based CSI prediction, in terms of mean UPT, gains are observed compared to both Benchmark#1 of the nearest historical CSI and Benchmark#2 of a non-AI/ML based CSI prediction approach:

- Compared to the benchmark of the nearest historical CSI:

- For FTP traffic:

- 4 sources observe 1.2%~4.9% gain;

- 2 sources observe 5.3%~10.58% gain;

- 2 sources observe 15.1% ~23.5% gain.

- 1 source observes loss of -1.3%~-13.8%.

- For full buffer traffic:

- 1 source observes 2%~3% gain;

- 2 sources observe 7.6%~15.6% gain.

- Compared to the benchmark of an auto-regression/Kalman filter based CSI prediction:

- For FTP traffic:

- 3 sources observe 0.7%~7.0% gain;

- 2 sources observe loss of -0.1%~-2.4%.

- 1 source observe loss of -3%~-17%.

- For full buffer traffic:

- 2 sources observes 0.6%~2.78% gain.

- 1 source observes 8.1%~11.5% gain.

The above results are based on the following assumptions:

- The same fixed UE speed of 30km/h or 60km/h is assumed for both training and inference

- The observation window considers to start as early as 15ms~50ms.

- A future 4ms or 5ms instance from the prediction output is considered for calculating the metric.

- Raw channel matrix is considered as model input

- The performance metric is mean UPT for Max rank 1.

- No post processing is considered.

- Note: Results refer to Table 5.28 of R1-2308344.

***5% UPT***

For the AI/ML based CSI prediction, in terms of 5% UPT, gains are observed compared to both Benchmark#1 of the nearest historical CSI and Benchmark#2 of a non-AI/ML based CSI prediction approach:

- Compared to the benchmark of the nearest historical CSI:

- For FTP traffic:

- 4 sources observe 1% ~9.7% gain;

- 5 sources observe 10%~26.4% gain;

- 1 source observes loss of -11.6%~-14%;

- For full buffer traffic:

- 3 sources observe 3.5%~35.3% gain;

- Compared to the benchmark of an auto-regression/Kalman filter based CSI prediction:

- For FTP traffic:

- 3 sources observe 0.18%~17.58% gain;

- 1 source observes -8.2%~-12.4% degradation;

- For full buffer traffic:

- 1 source observes 6.7% ~15.4% gain.

- 1 source observes -2% degradation

The above results are based on the following assumptions:

- The same fixed UE speed of 30km/h or 60km/h is assumed for both training and inference

- The observation window considers to start as early as 15ms~50ms.

- A future 4ms or 5ms instance from the prediction output is considered for calculating the metric.

- Raw channel matrix is considered as model input

- The performance metric is 5% UPT for Max rank 1.

- No post processing is considered.

- Note: Results refer to Table 5.29 of R1-2308344.

#### 6.2.2.7 Generalization evaluations for CSI prediction

***Generalization over UE speeds***

For the *generalization verification* of AI/ML based CSI prediction *over various UE speeds* compared to the generalization Case 1 where the AI/ML model is trained with dataset subject to a certain UE speed#B and applied for inference with a same UE speed#B,

- For generalization Case 2, generalized performance may be achieved for certain combinations of UE speed#A and UE speed#B but not for others:

- If UE speed#B is 10 km/h & UE speed#A is 30 km/h, 2 sources observe a generalized performance of less than -1.4% degradation.

- Note: 1 company still observes significant degradation (-11.3%~-13.4% loss).

- If UE speed#B is either 30 km/h or 60 km/h or 120 km/h, or if UE speed#B is 10km/h and UE speed#A is either 60km/h or 120km/h, 11 sources observe that moderate/significant performance degradations are suffered:

- For UE speed#B is 10 km/h & UE speed#A is either 60 km/h or 120 km/h, 1 source observes moderate degradation (-2.3% loss), 3 sources observe significant degradation (-5.5%~-61% loss).

- For UE speed#B is 30 km/h & UE speed#A is either 10 km/h, 60 km/h or 120 km/h, 2 sources observe moderate degradation (-2.01%~-4.62% loss), 9 sources observe significant degradation (-5%~-72.37% loss).

- For UE speed#B is 60 km/h & UE speed#A is either 10 km/h, 30 km/h or 120 km/h, 1 source observes moderate degradation (-3% loss), 10 sources observe significant degradation (-7.8%~-76.85% loss).

- For UE speed#B is 120 km/h & UE speed#A is either 30 km/h or 60 km/h, 1 source observes moderate degradation (-3.4% loss), 5 sources observe significant degradation (-7.55%~-56.3% loss).

- For generalization Case 3, generalized performance of the AI/ML model can be achieved in general (0%~-4.45% loss) for UE speed#B subject to any of 10 km/h, 30 km/h, 60 km/h and 120 km/h, if the training dataset is constructed with data samples subject to multiple UE speeds including UE speed#B, as observed by 11 sources.

- For UE speed#B is 10 km/h, minor loss (-0.2%~-1.7%) are observed by 4 sources.

- For UE speed#B is 30 km/h, minor loss (-0.2%~-1.34%) or positive gain are observed by 5 sources, moderate loss (-4.07%~-4.2%) are observed by 2 sources.

- For UE speed#B is 60 km/h, minor loss (-0.05%~-2%) are observed by 4 sources, moderate loss (-3.76%~-4.65%) are observed by 2 sources.

- For UE speed#B is 120 km/h, moderate loss (-2%~-4.45%) are observed by 4 sources.

- Note: For generalization Case 3, 6 sources observe significant performance degradations (-5%~-43.6% loss) for UE speed#B subject to 10 km/h, 30 km/h, 60 km/h, but compared with generalization Case 2, in general the performance is still improved.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Raw channel matrix is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- The performance metric is SGCS in linear value for layer 1/2/3/4.

- No spatial consistency is considered.

- Note: Results refer to Table 5.5 of R1-2308340.

#### 6.2.2.8 Summary of Performance Results for CSI feedback enhancement

The following aspects have been studied for the evaluation on AI/ML based **CSI compression** in Rel-18:

- From the perspective of basic performance gain over non-AI/ML benchmark (assuming 1 on 1 joint training without considering generalization),

o It has been studied with corresponding observations on:

 the metrics of SGCS, mean UPT, 5% UPT, CSI feedback overhead reduction

 the benchmark of R16 Type II codebook

o It has been studied but is lack of observations on:

 the metric of NMSE

 the benchmarks of Type I codebook and R17 Type II codebook

o It has been studied with corresponding observations on complexity but without comparison with non-AI/ML.

- From the perspective of AI/ML solutions (assuming 1 on 1 joint training without considering generalization),

o It has been studied with corresponding observations on: model input/output type, monitoring for intermediate KPI (including NW side monitoring and UE side monitoring), quantization methods (including quantization awareness for training, and quantization format), and high resolution ground-truth CSI for training, with the metric of SGCS.

o It has been studied but is lack of observations on: the options of CQI/RI calculation, and the options of rank>1 solution

- From the perspective of generalization over various scenarios (assuming 1 on 1 joint training),

o It has been studied with corresponding observations on (with the metric of SGCS):

 the scenarios including various deployment scenarios, various outdoor/indoor UE distributions, various carrier frequencies, and various TxRU mappings

 the approach of dataset mixing (generalization Case 3)

o It has been studied but is lack of observations on:

 other aspects of scenarios

 the approach of fine-tuning

- From the perspective of scalability over various configurations (assuming 1 on 1 joint training),

o It has been studied with corresponding observations on (with the metric of SGCS):

 the configurations including various bandwidths/frequency granularities, various CSI feedback payloads, and various antenna port numbers

 the approach of dataset mixing (generalization Case 3), and the approach of fine-tuning for CSI feedback payloads

 the scalability solutions

o It has been studied but is lack of observations on:

 other aspects of configurations

 the approach of fine-tuning for configurations other than CSI feedback payloads

- From the perspective of multi-vendor joint training (without considering generalization),

o It has been studied with corresponding observations on (with the metric of SGCS):

 joint training between 1 NW part model and M>1 UE part models, and joint training between 1 UE part model and N>1 NW part models

o It has been studied but is lack of observations on:

 joint training between N>1 NW part models and M>1 UE part models

 performance comparison between simultaneous training and sequential training

- From the perspective of separate training (without considering generalization),

o It has been studied with corresponding observations on (with the metric of SGCS):

 NW first training, including 1 NW part model to 1 UE part model with same backbone and with different backbones, and 1 UE part model to N>1 NW part models

 UE first training, including 1 NW part model to 1 UE part model with same backbone and with different backbones, and 1 NW part model to M>1 UE part models

 Impact of shared dataset under 1 NW part model to 1 UE part model for NW first training and UE first training

o It has been studied but is lack of observations on:

 the metric of air-interface overhead of information (e.g., dataset) sharing

The following aspects have been studied for the evaluation on AI/ML based **CSI prediction**:

- From the perspective of basic performance gain over non-AI/ML benchmark (without considering generalization),

o It has been studied with corresponding observations on:

 the metrics of SGCS, mean UPT, 5% UPT;

 the benchmarks of nearest historical CSI and auto-regression/Kalman filter based CSI prediction.

 Note: the benchmark of level x based CSI prediction is represented by generalization cases.

o It has been studied but is lack of observations on:

 the impact of modeling spatial consistency

 the metrics of NMSE

o It has been studied with corresponding observations on complexity but without comparison with non-AI/ML

- From the perspective of AI/ML solutions (without considering generalization),

o It has been studied with corresponding observations on (with the metric of SGCS and the benchmark of nearest historical CSI): impact of input type, impact of UE speed, impact of prediction window, impact of observation window

- From the perspective of generalization over various scenarios,

o It has been studied with corresponding observations on (with the metric of SGCS):

 the scenario including various UE speeds

 the approach of dataset mixing (generalization Case 3)

o It has been studied but is lack of observations on:

 various deployment scenarios, various carrier frequencies, and other aspects of scenarios.

 the approach of fine-tuning

- From the perspective of scalability over various configurations, it has been studied but is lack of observations.

Based on the evaluation for **CSI compression**, the following high-level observations are provided:

- From the perspective of basic performance gain over non-AI/ML benchmark, AI/ML based CSI compression outperforms Rel-16 eType II CB in general under 1-on-1 joint training and generalization Case 1, where

o 0.2%~2%/-0.3%~6%/-4%~6% gains of mean UPT as shown in Figure 6.2.2.8-1 through Figure 6.2.2.8-3 are observed for Max rank 1/2/4, respectively, under RU≤39%.

o 0.1%~4%/-0.5%~10%/-1.8%~12.22% gains of mean UPT as shown in Figure 6.2.2.8-4 through Figure 6.2.2.8-6 are observed for Max rank 1/2/4, respectively, under RU40%-69%.

o 0.23%~9%/-0.2%~15%/-1%~17% gains of mean UPT as shown in Figure 6.2.2.8-7 through Figure 6.2.2.8-9 are observed for Max rank 1/2/4, respectively, under RU≥70%.



Figure 6.2.2.8-1: Mean UPT gain, Max Rank 1 (RU≤39%), x-axis means index of source



Figure 6.2.2.8-2: Mean UPT gain, Max Rank 2 (RU≤39%), x-axis means index of source



Figure 6.2.2.8-3: Mean UPT gain, Max Rank 4 (RU≤39%), x-axis means index of source



Figure 6.2.2.8-4: Mean UPT gain, Max Rank 1 (RU40%-69%), x-axis means index of source



Figure 6.2.2.8-5: Mean UPT gain, Max Rank 2 (RU40%-69%), x-axis means index of source



Figure 6.2.2.8-6: Mean UPT gain, Max Rank 4 (RU40%-69%), x-axis means index of source



Figure 6.2.2.8-7: Mean UPT gain, Max Rank 1 (RU≥70%), x-axis means index of source



Figure 6.2.2.8-8: Mean UPT gain, Max Rank 2 (RU≥70%), x-axis means index of source



Figure 6.2.2.8-9: Mean UPT gain, Max Rank 4 (RU≥70%), x-axis means index of source

- From the perspective of CSI feedback overhead reduction over non-AI/ML, AI/ML based CSI compression achieves CSI feedback reduction compared with Rel-16 eType II CB in general under 1-on-1 joint training and generalization Case 1, where 4 sources observe the CSI feedback overhead reduction of 10.24%~60%/10%~58.33%/8%~79% for Max rank 1/2/4, respectively, under FTP traffic.

- From the perspective of AI/ML complexity, a majority of 25 sources adopt the CSI generation model subject to the computational complexity in units of FLOPs from 10M to 800M, and 26 sources adopt the CSI reconstruction model subject to the FLOPs from 10M to 1100M. The actual model complexity may differ from the model complexity in the evaluation with respect to platform-dependent optimization on model implementations. In addition, the complexity between AI/ML and non-AI/ML benchmark is not compared.

- From the perspective of model input/output type, it is more beneficial by considering precoding matrix as the model input (for CSI generation part)/output (for CSI reconstruction part) than explicit channel matrix.

- From the perspective of intermediate KPI based monitoring,

o For the monitoring at NW side, increased monitoring accuracy can be achieved by considering R16 eType II CB with new/larger parameter(s) as the ground-truth CSI format for monitoring. On the other hand, the new/larger parameter(s) would lead to increased air-interface overhead compared to R16 eType II CB with legacy parameters.

o For the monitoring at UE side, performance can be monitored with smaller air-interface overhead by considering proxy model at UE compared with monitoring at NW side. On the other hand, the monitoring accuracy may be impacted by the design/robustness of the proxy model.

o Note: the complexity aspect for Case 1, Case 2-1 and Case 2-2 is not evaluated.

- From the perspective of quantization methods for CSI feedback,

o For the quantization awareness for training, it is beneficial to consider quantization aware training with fixed/pre-configured quantization method/parameters (Case 2-1) or jointly updated quantization method/parameters (Case 2-2) to avoid severe performance degradation. In particular, it is more beneficial in performance for Case 2-2 over Case 2-1 under vector quantization format (VQ).

o For the quantization format, VQ format achieves comparable performance with scalar quantization format (SQ) in general, where VQ achieves better performance than SQ in some cases while worse in some other cases.

- From the perspective of high resolution ground-truth CSI for training, compared to unquantized ground-truth CSI (e.g., Float32), taking R16 eType II CB with new/larger parameter(s) as the ground-truth CSI format for training data collection can achieve significant overhead reduction without causing severe performance degradation; taking scalar quantization format for training data collection can achieve moderate overhead reduction without causing severe performance degradation. On the other hand, the R16 eType II CB with new/larger parameter(s) would lead to increased overhead compared to R16 eType II CB with legacy parameters

o For ground-truth CSI format, 5 sources observe R16 eType II CB with new/larger parameter(s) outperforms R16 eType II CB with legacy parameter, while one source observes R16 eType II CB with legacy parameter is already close to Float32 with particular dataset processing technique.

o Note: the complexity aspect is not evaluated.

- From the perspective of generalization over scenarios, or scalability over configurations that have been evaluated, compared to generalization Case 1 where the AI/ML model is trained with dataset subject to a certain scenario#B/configuration#B and applied for inference with a same scenario#B/configuration#B,

o For generalization Case 2 where the AI/ML model is trained with dataset from a different scenario#A/configuration#A, generalized performance may be achieved for some certain combinations of scenario#A/configuration#A and scenario#B/configuration#B but not for others.

o For generalization Case 3 where the training dataset is constructed with data samples subject to more than one scenario/configuration (evaluations studied up to four scenarios/configurations) including scenario#B/configuration#B, generalized performance of the AI/ML model can be achieved.

o In particular, appropriate scalability solution (e.g., truncation/padding, adaptive quantization granularities, adaptation layer in the AI/ML model) may need to be performed to scale the dimensions of the AI/ML model when the training dataset includes data samples subject to configuration#A which has different input/output dimension than configuration#B.

- From the perspective of training collaboration types, compared to 1-on-1 joint training, both multi-vendor joint training and separate training with procedures given in clause 6.2.1 may suffer performance loss.

o In particular, for multi-vendor joint training, minor or moderate degradation is observed.

o In particular, for separate training with procedure given in clause 6.2.1, the performance loss depends on the factors such as backbone alignment, and multi-vendor training behavior:

 For separate training of 1 NW part model and 1 UE part model, under both NW first training and UE first training, if backbones are aligned between the two sides, minor degradation is observed; otherwise, additional degradation is observed, leading to minor or moderate performance degradation.

 For NW first training with 1 UE part model to N>1 NW part models, or UE first training with 1 NW part model to M>1 UE part models, additional degradation is observed, leading to minor, moderate, or significant performance degradation, depending on the training approach.

 As a note, other procedures of separate training are not extensively evaluated.

Based on the evaluation for **CSI prediction**, the following high-level observations are provided:

- From the perspective of basic performance gain over non-AI/ML benchmark, under the same UE speed for training and inference,

o AI/ML based CSI prediction outperforms the benchmark of the nearest historical CSI in general, where the majority of sources observe up to 10.6% gain in terms of mean UPT.

o for AI/ML based CSI prediction over non-AI/ML based CSI prediction, 3 sources observe 0.7%~7% gain while 2 sources observe performance loss of -0.1%~-17% in terms of mean UPT.

- From the perspective of AI/ML complexity, a majority of 16 sources adopt the model subject to the computational complexity in units of FLOPs from 0.1M to 1000M. The actual model complexity may differ from the model complexity in the evaluation with respect to platform-dependent optimization on model implementations. In addition, the complexity between AI/ML and non-AI/ML benchmark is not compared.

- From the perspective of model input/output type, it is more beneficial in performance by considering raw channel matrix as the model input than precoding matrix

- From the perspective of AI/ML solutions, the gain of AI/ML based CSI prediction over the benchmark of the nearest historical CSI is impacted by the observation window length, prediction window length, and UE speed

- From the perspective of generalization over UE speeds that have been evaluated, compared to generalization Case 1 where the AI/ML model is trained with dataset subject to a certain UE speed#B and applied for inference with a same UE speed#B,

o For generalization Case 2 where the AI/ML model is trained with dataset from a different UE speed#A, generalized performance may be achieved for some certain combinations of UE speed#A and UE speed#B but not for others

o For generalization Case 3 where the training dataset is constructed with data samples subject to multiple UE speeds including UE speed#B, generalized performance of the AI/ML model can be achieved in general

## 6.3 Beam management

### 6.3.1 Evaluation assumptions, methodology and KPIs

Figure 6.3.1-1 provides an example for the inference procedure for beam management for BM-Case1 and BM-Case2. Measurements based on Set B of beams are used as model input. In addition, beam ID information may be also provided as input to the AI/ML model. Based on model output (e.g., probability of each beam in Set A to be the Top-1 beam, predicted L1-RSRPs), Top-1/N beam(s) among Set A of beams can be predicted and/or potentially with predicted L1-RSRPs (depending on the labelling). In the evaluation, for BM-Case 1, the measurements of Set B (otherwise stated) are used as model input to predict Top-1/N beams from Set A, and for BM-Case2, the measurements from historic time instance(s) are used as model input for temporal DL beam prediction of beams from Set A. In the evaluation, the cases that Set A and Set B are different (Set B is NOT a subset of Set A), and Set B is a subset of Set A for both BM-Case1 and BM-Case2, and case that Set A and Set B are the same for BM-Case2 are considered. And the performance of DL Tx beam prediction and DL Tx-Rx beam pair prediction is evaluated.

For both BM-Case1 and BM-Case2, UE can report the prediction result to NW based on the output of a UE-side model, or NW can predict the Top-1/N beam(s) based on the reported measurements of Set B for a NW-side model.



Figure 6.3.1-1: An example of the inference procedure for beam management.

For dataset construction and performance evaluation (if applicable) in the AI/ML for beam management use case, *system level simulation* approach is adopted as baseline. *Link level simulation* is optionally adopted.

***KPIs*:**

- Model complexity and computational complexity.

Beam prediction accuracy related KPIs, including:

**- Top-1 genie-aided Tx beam** considers the following definitions:

- Option A (baseline), the Top-1 genie-aided Tx beam is the Tx beam that results in the largest L1-RSRP over all Tx and Rx beams

- Option B (optional), the Top-1 genie-aided Tx beam is the Tx beam that results in the largest L1-RSRP over all Tx beams with specific Rx beam(s)

- Specific Rx beam(s) are to be reported. Note: specific Rx beams are a subset of all Rx beams.

**- Top-1 genie-aided Tx-Rx beam pair** considers the following definitions:

- Option A: The Tx-Rx beam pair that results in the largest L1-RSRP over all Tx and Rx beams

- Other options not precluded and can be reported

- Average L1-RSRP difference of Top-1 predicted beam:

- The difference between the ideal L1-RSRP of Top-1 predicted beam and the ideal L1-RSRP of the Top-1 genie-aided beam

- Beam prediction accuracy (%):

- Top-1 (%): the percentage of "the Top-1 genie-aided beam is Top-1 predicted beam"

- Top-K/1 (%): the percentage of "the Top-1 genie-aided beam is one of the Top-K predicted beams"

- Top-1/K (%) (Optional): the percentage of "the Top-1 predicted beam is one of the Top-K genie-aided beams"

- Where K >1 and values can be reported

- CDF of L1-RSRP difference for Top-1 predicted beam

- Beam prediction accuracy (%) with 1dB margin for Top-1 beam

- The beam prediction accuracy (%) with 1dB margin is the percentage of the Top-1 predicted beam "whose ideal L1-RSRP is within 1dB of the ideal L1-RSRP of the Top-1 genie-aided beam"

- Other beam prediction accuracy related KPIs are not precluded and can be reported

Impact of quantization error of imputed L1-RSRP (for training and inference) is to be studied. Existing quantization granularity of L1-RSRP (i.e., 1dB for the best beam, 2dB for the difference to the best beam) is the starting point for evaluation at least for NW-sided model.

The performance impact of the relative L1-RSRP measurement error can be optionally evaluated for both DL Tx beam and beam pair prediction, where the relative L1-RSRP measurement error can be modelled as noise among beams as a starting point:

- Additive Gaussian noise with 95% of the density function within the measurement accuracy range, and/or uniformly distributed noise for the error due to baseband and/or RF impairment.

- Other modelling methods are not precluded and can be reported by companies.

- Companies’ report includes how to model the measurement error and the measurement accuracy range in training and test data and labels.

- Companies’ report includes the baseline performance with the relative L1-RSRP measurement error

System performance related KPIs, including:

- UE throughput: CDF of UE throughput, average and 5%-ile UE throughput

- RS overhead reduction for BM-Case1:

- Option 1: "RS " OH reduction[%]=1-N/M

- where N is the number of beams (pairs) (with reference signal (SSB and/or CSI-RS)) required for measurement for AI/ML

- where M is the total number of beams (pairs) to be predicted

- Option 2: "RS " OH reduction[%]=1-N/M

- where N is the total number of beams (pairs) (with reference signal (SSB and/or CSI-RS)) required for measurement for AI/ML, including the beams (pairs) required for additional measurements before/after the prediction if applicable

- where M is the total number of beams (pairs) (with reference signal (SSB and/or CSI-RS)) required for measurement for baseline scheme, including the beams (pairs) required for additional measurements before/after the prediction if applicable

- Companies report the assumption on additional measurements

- RS overhead reduction for BM-Case2, when Top-1 and Top-K beam (pairs) are inferred:

- "RS " OH reduction[%]=1-N/M

- where N is the total number of beams (pairs) (with reference signal (SSB and/or CSI-RS)) required for measurement for AI/ML, including the beams (pairs) required for additional measurements before/after the prediction if applicable.

- where M is the total number of beams (pairs) (with reference signal (SSB and/or CSI-RS)) required for measurement for baseline scheme

- Companies report the assumption on additional measurements.

- Companies report the assumption on baseline scheme.

- Companies report the assumption on T1 and T2.

- Other System performance related KPIs are not precluded and can be reported by companies

To calculate the measurement/RS overhead reduction and summarize results for BM-Case 2, at least when Top-1 beam (pair) is inferred:

**- Case A:** based on number of measurements/RSs and prediction time. An example is shown in Figure 6.3.1-2.

- where T2 is the time duration for beam prediction

- where Mt is the number of time instances for measurement as AI/ML inputs with a periodicity of Tper

- where Pt is the number of time instance(s) for prediction with a periodicity of Tper in T2

**- In this case,** the non-AI baseline is Option 1 (measured all the beams at each time instance(s) for prediction with a periodicity of Tper in T2)

- For Set B= Set A, the RS overhead reduction is 1-Mt/(Mt+Pt).

- For Set B (N beams, same number in each time instance) is a subset of Set A (M beams), the RS overhead reduction is

- N\*Mt/(M\*(Mt+Pt)) if no sliding window

- 1-N/M if considering sliding window

**- Case B:** based on a periodicity T of the required reference signals for measurements to achieve a certain beam prediction accuracy. An example is shown in Figure 6.3.1-3.

- For non-AI baseline (Option 2), every T=X ms reference signals for measurements are needed

- For AI, every T=Y ms, reference signals for measurements are needed

**- In this case,**

- For Set B = Set A, the RS overhead reduction is 1-X/Y.

- For Set B (N beams) is a subset of Set A (M beams), the RS overhead reduction is [1-XN/(YM)].

**- Case B+:** based on Y times of a given minimal periodicity Tper of the reference signals for measurements. An example is shown in Figure 6.3.1-4.

- For non-AI baseline (Option 1), UE measures all the reference signals of Set A every Tper

- For AI, UE measures the reference signals of Set B every Y times of Tper

- In this case, prediction time is defined as the time from each measurement instance to the latest prediction instance before the next measurement instance.

**- In this case,** the non-AI baseline is Option 1 (measured all the beams at each time instance(s) for prediction with a periodicity of Tper, which is reported by companies)

- For Set B= Set A, the RS overhead reduction is 1-1/Y.

- For Set B (N beams) is a subset of Set A (M beams), the RS overhead reduction is 1-N/(YM).



Figure 6.3.1-2: Example for Case A



Figure 6.3.1-3: Example for Case B



Figure 6.3.1-4: Example for Case B+

For both BM-Case1 and BM-Case2 when Set B is a subset of or different from Set A, a certain RS/measurement overhead is assumed to summarize the evaluation results for Top-1(%) beam prediction accuracy. With additional measurements among predicted Top-K beam (pairs) (i.e., with additional RS/measurement overhead), Top-1 beam (pair) can be obtained by finding a best beam (pair) among the K predicted beams (pairs) with the beam prediction accuracy of Top-K/1(%) if no genie-aid Top-1 beam change out of the K predicted beam (pairs) during the additional measurements.

Note: This is to explain the potential implications and relations of Top-1(%) and Top-K/1(%) beam prediction accuracy metrics defined in evaluations agenda item with regards to RS overhead and additional measurement. The corresponding specification impact is a separate discussion.

Other KPIs, including:

- UCI report overhead (e.g., number of UCI reports and UCI payload size) and/or UCI overhead reduction for inference of AI/ML model can be reported, at least for NW side beam prediction

- UCI overhead reduction = 1- Total UCI payload size for AI/ML/Total UCI payload size of baseline.

- Companies expected to report detailed assumption of UCI for AI/ML and baseline, e.g., including quantization mechanism.

- Latency reduction:

- (1 – (Total transmission time of N beams) / (Total transmission time of M beams))

- where N is the number of beams (with reference signal (SSB and/or CSI-RS)) in the input beam set required for measurement

- where M is the total number of beams

- Power consumption reduction

For AI/ML models, which provide L1-RSRP as the model output, the accuracy of predicted L1-RSRP is to be evaluated. Companies optionally report average (absolute value)/CDF of the predicted L1-RSRP difference, where the predicted L1-RSRP difference is defined as the difference between the predicted L1-RSRP of Top-1 predicted beam and the ideal L1-RSRP of the same beam.

***Model generalization*:**

In the context of model generalization, scenarios may mean various deployment scenarios, various outdoor/indoor UE distributions, various UE mobility assumptions. Similarly, configurations may mean various UE parameters, various gNB settings, Various Set B of beam(pairs). The selected scenarios/configurations for generalization verification may consider the AI model inference node (e.g., @UE or @gNB) and use case (e.g., BM-Case1, or BM-Case2). Specifically, the following generalizations could be considered and clause 6.3.2 presents those which have been actually simulated by companies:

- Scenarios:

- Various deployment scenarios, e.g., UMa, UMi and others; e.g., 200m ISD or 500m ISD and others; e.g., same deployment, different cells with different configuration/assumption; e.g., gNB height and UE height;

- Various outdoor/indoor UE distributions, e.g., 100%/0%, 20%/80%, and others

- Various UE mobility, e.g., 3km/h, 30km/h, 60km/h and others

- Configurations (parameters and settings):

- Various UE parameters, e.g., number of UE Rx beams (including number of panels and UE antenna array dimensions)

- Various gNB settings, e.g., DL Tx beam codebook (including various Set A of beam(pairs) and gNB antenna array dimensions)

- Various Set B of beam (pairs)

- T1 for measurement /T2 for prediction for BM-Case2

- Other scenarios/configurations(parameters and settings) are not precluded and can be reported

Companies to report the selected scenarios/configurations for generalization verification. Note: other approaches for achieving good generalization performance for AI/ML-based schemes are not precluded.

The following cases are considered for verifying the *generalization performance* of an AI/ML model over various scenarios/configurations as a starting point:

- **Case 1**: The AI/ML model is trained based on training dataset from one Scenario#A/Configuration#A, and then the AI/ML model performs inference/test on a dataset from the same Scenario#A/Configuration#A

**- Case 2**: The AI/ML model is trained based on training dataset from one Scenario#A/Configuration#A, and then the AI/ML model performs inference/test on a different dataset than Scenario#A/Configuration#A, e.g., Scenario#B/Configuration#B, Scenario#A/Configuration#B

**- Case 3**: The AI/ML model is trained based on training dataset constructed by mixing datasets from multiple scenarios/configurations including Scenario#A/Configuration#A and a different dataset than Scenario#A/Configuration#A, e.g., Scenario#B/Configuration#B, Scenario#A/Configuration#B, and then the AI/ML model performs inference/test on a dataset from a single Scenario/Configuration from the multiple scenarios/configurations, e.g., Scenario#A/Configuration#A, Scenario#B/Configuration#B, Scenario#A/Configuration#B.

- Notes: Companies to report the ratio for dataset mixing. Number of the multiple scenarios/configurations can be larger than two.

- The following case for generalization verification, can be optionally considered by companies:

- **Case 2A**: The AI/ML model is trained based on training dataset from one Scenario#A/Configuration#A, and then the AI/ML model is updated based on a fine-tuning dataset different than Scenario#A/Configuration#A, e.g., Scenario#B/Configuration#B, Scenario#A/Configuration#B. After that, the AI/ML model is tested on a different dataset than Scenario#A/Configuration#A, e.g., subject to Scenario#B/Configuration#B, Scenario#A/Configuration#B.

- Companies to report the fine-tuning dataset setting (e.g., size of dataset) and the improvement of performance.

Further details on evaluation assumptions

The following options are studied on the selection of Set B of beams (pairs):

- **Option 1**: Set B is fixed across training and inference

**- Option 2**: Set B is variable (e.g., different beams (pairs) patterns in each time instance/report/measurement during training and/or inference)

**- Opt 2A**: Set B is changed following a set of pre-configured patterns

**- Opt 2B**: Set B is randomly changed among pre-configured patterns

**- Opt 2C**: Set B is randomly changed among Set A beams (pairs)

**- Opt 2D**: Set B is a subset of measured beams (pairs) Set C (including Set B = Set C), e.g. Top-K beams(pairs) of Set C

- The number of beams(pairs) in Set B can be fixed or variable

- Companies report the number of pre-configured patterns used in the evaluation for Option 2: Set B is variable if applicable (e.g. Opt A and Opt B)

- Note: BM-Case1 and BM-Case2 may be considered for different option.

- Note: This does not preclude the alternative that Set B is different from Set A.

For the evaluation of Option 2: Set B is variable (e.g., different beams (pairs) patterns in each time instance/report/measurement during training and/or inference), study the following options as AI/ML model inputs:

- Alt 1: *Implicit* information of Tx beam ID and/or Rx beam ID

- e.g., measurements of Set B of beams together with default values (e.g., 0) for the beams not in Set B are used as AI inputs in a certain order/ matrix/ vector. Detailed assumption can be reported.

- Alt 2: Tx beam ID and/or Rx beam ID is used as inputs of AI/ML *explicitly*.

For the purpose of DL Tx beam prediction evaluations, consider the following options for Rx beam as AI/ML model input for training and/or inference if applicable:

- Option 1: Measurements of the "best" Rx beam with exhaustive beam sweeping for each model input sample.

- Companies expected to report how to select the "best" Rx beam(s).

- Option 2: Measurements of specific Rx beam(s).

- Companies expected to report how to select specific Rx beam(s).

- Option 3: Measurements of random Rx beam(s) per model input sample.

- Option 4: Measurements of quasi-optimal Rx beam (i.e., not all the measurements as inputs of AI/ML are from the "best" Rx beam) with less measurement/RS overhead compared to exhaustive Rx beam sweeping.

- Identify the quasi-optimal Rx beams to be utilized for measuring Set B/Set C based on the previous measurements. Companies can report the time information and beam type (e.g., whether the same Tx beam(s) in Set B) of the reference signal to use. Companies expected to report the measurement/RS overhead together with the beam prediction accuracy, as well as, how to find the quasi-optimal Rx beam with "previous measurement".

- Other options are not precluded and can be reported by companies.

Performance with different types of labels are studied considering the following:

- Option 1a: Top-1 beam(pair) in Set A

- Option 1b: Top-K beam (pair)s in Set A

- Option 2a: L1-RSRPs per beam of all the beams(pairs) in Set A

- Option 2b: Top-K beam(pair)s in Set A and the corresponding L1-RSRPs

- Option 2c: Top-1 beam(pair) in Set A and the corresponding L1-RSRP

***Evaluation assumptions:***

Table 6.3.1-1 presents the baseline system level simulation assumptions for AI/ML in beam management evaluations.

Table 6.3.1-1: Baseline System Level Simulation assumptions for AI/ML in beam management evaluations

|  |  |
| --- | --- |
| Parameter | Value |
| Frequency Range | FR2 @ 30 GHz; SCS: 120 kHz |
| Deployment | 200m ISD, 2-tier model with wrap-around (7 sites, 3 sectors/cells per site)Other deployment assumption is not precluded |
| Channel model | UMa with distance-dependent LoS probability function defined in Table 7.4.2-1 in TR 38.901. |
| System BW | 80MHz |
| UE Speed | For spatial domain beam prediction: 3km/hFor time domain beam prediction: 30km/h (baseline), 60km/h (optional) 90km/h (optional), 120km/h (optional)Other values are not precluded |
| UE distribution | 10 UEs per sector/cell for system performance related KPI (if supported) [e.g., throughput] for full buffer traffic (if supported) evaluation (model inference).X UEs per sector/cell for system performance related KPI for FTP traffic (if supported) evaluation (model inference).Other values are not precluded. Number of UEs per sector/cell during data collection (training/testing) is reported by companies if relevant.For spatial domain beam prediction (optional to compare different UE distributions assumptions):- Option 1: 80% indoor ,20% outdoor as in TR 38.901- Option 2: 100% outdoorFor time domain prediction: 100% outdoor |
| Transmission Power | Maximum Power and Maximum EIRP for base station and UE as given by corresponding scenario in 38.802 (Table A.2.1-1 and Table A.2.1-2) |
| BS Antenna Configuration | Antenna setup and port layouts at gNB: (4, 8, 2, 1, 1, 1, 1), (dV, dH) = (0.5, 0.5) λOther assumptions are not precluded. Companies to explain TXRU weights mapping.Companies to explain beam selection.Number of BS beams: 32 or 64 downlink Tx beams (max number of available beams) at NW side. Other values, e.g., 256 not precluded. |
| BS Antenna radiation pattern | TR 38.802 Table A.2.1-6, Table A.2.1-7 |
| UE Antenna Configuration | Antenna setup and port layouts at UE: (1, 4, 2, 1, 2, 1, 1), 2 panels (left, right)Other assumptions are not precludedCompanies to explain TXRU weights mapping.Companies to explain beam and panel selection.Number of UE beams: 4 or 8 downlink Rx beams (max number of available beams) per UE panel at UE side. Other values, e.g., 16 not precluded. |
| UE Antenna radiation pattern | TR 38.802 Table A.2.1-8, Table A.2.1-10 |
| Beam correspondence | Companies to explain beam correspondence assumptions (in accordance to the two types agreed in RAN4) |
| Link adaptation | Based on CSI-RS |
| Traffic Model | For system performance related KPI (if supported) evaluation (model inference), companies report either of the following traffic model: Option 1: Full buffer Option 2: FTP model with detail assumptions (e.g., FTP model 1, FTP model 3) |
| Inter-panel calibration for UE | Ideal, non-ideal following 38.802 (optional) – Explain any errors |
| Control and RS overhead | Companies report details of the assumptions |
| Control channel decoding | Ideal or Non-ideal (Companies explain how it is modelled) |
| UE receiver type | MMSE-IRC as the baseline, other advanced receiver is not precluded |
| BF scheme | Companies to explain what scheme is used |
| Transmission scheme | Multi-antenna port transmission schemesNote: Companies explain details of the using transmission scheme. |
| Other simulation assumptions | Companies to explain serving TRP selectionCompanies to explain scheduling algorithm |
| Other potential impairments | Not modelled (assumed ideal).If impairments are included, companies will report the details of the assumed impairments |
| BS Tx Power | 40 dBm (baseline)Other values (e.g., 34 dBm) not precluded |
| Maximum UE Tx Power | 23 dBm |
| BS receiver Noise Figure | 7 dB |
| UE receiver Noise Figure | 10 dB |
| Inter site distance | 200 m |
| BS Antenna height | 25 m |
| UE Antenna height | 1.5 m |
| Car penetration Loss | 38.901, sec 7.4.3.2: μ = 9 dB, σp = 5 dB |
| UE measurements/reports | At least for Temporal Downlink beam prediction: - Periodicity of time instance for each measurement/report in T1: 20ms, 40ms, 80ms, [100ms], 160ms, [960ms]. Other values can be reported.- Number of time instances for measurement/report in T1 can be reported. Time instance(s) for prediction can be reported. |
| Scenario | Dense Urban (macro-layer only, TR 38.913) is the basic scenario for dataset generation and performance evaluation. Other scenarios are not precluded.  |
| Spatial consistency  | At least for BM-Case1, companies report the one of spatial consistency procedures: - Procedure A in TR38.901- Procedure B in TR38.901 |
| UE trajectory model | UE trajectory model is defined at least for *temporal beam prediction* in initial phase of the evaluation. Further details below. UE trajectory model is not necessarily to be defined at least for *spatial-domain beam prediction* in initial phase of the evaluation. |
| UE rotation | UE speed to be reported. Note: UE rotation speed = 0, i.e., no UE rotation, is not precluded |
| Baseline for performance evaluation | For *temporal beam prediction*: - Option 1: Select the best beam for T2 within Set A of beams based on the measurements of all the RS resources or all possible beams from Set A of beams at the time instants within T2 - Option 2: Select the best beam for T2 within Set A of beams based on the measurements of all the RS resources from Set B of beams at the time instants within T1- Companies to explain the detail on how to select the best beam for T2 from Set A based on the measurements in T1.where T2 is the time duration for the best beam selection, and T1 is a time duration to obtain the measurements of all the RS resource from Set B of beams. T1 and T2 are aligned with those for AI/ML based methods. Whether Set A and Set B are the same or different depend on the sub-use case. Other options are not precluded.For *spatial-domain beam prediction*: - Option 1: Select the best beam within Set A of beams based on the measurement of all RS resources or all possible beams of beam Set A (exhaustive beam sweeping) - Option 2: Select the best beam within Set A of beams based on the measurement of RS resources from Set B of beams- Other options are not precluded. |

For temporal beam prediction, the following options are considered as a starting point for *UE trajectory model*. Companies report further changes or modifications from those. Other options are not precluded. UE orientation can be independently modelled from UE moving trajectory. Other UE orientation model is not precluded:

- Option 1: Linear trajectory model with random direction change.

- UE moving trajectory: UE will move straight along the selected direction to the end of an time interval, where the length of the time interval is provided by using an exponential distribution with average interval length, e.g., 5s, with granularity of 100 ms.

- UE moving direction change: At the end of the time interval, UE will change the moving direction with the angle difference A\_diff from the beginning of the time interval, provided by using a uniform distribution within [-45°, 45°].

- UE moves straight within the time interval with the fixed speed.

- Option 2: Linear trajectory model with random and smooth direction change.

- UE moving trajectory: UE will change the moving direction by multiple steps within an time internal, where the length of the time interval is provided by using an exponential distribution with average interval length, e.g., 5s, with granularity of 100 ms.

- UE moving direction change: At the end of the time interval, UE will change the moving direction with the angle difference A\_diff from the beginning of the time interval, provided by using a uniform distribution within [-45°, 45°].

- The time interval is further broken into N sub-intervals, e.g. 100ms per sub-interval, and at the end of each sub-interval, UE change the direction by the angle of A\_diff/N.

- UE moves straight within the time sub-interval with the fixed speed.

- Option 3: Random direction straight-line trajectories.

- Initial UE location, moving direction and speed: UE is randomly dropped in a cell, and an initial moving direction is randomly selected, with a fixed speed.

- The initial UE location should be randomly drop within the following blue area:



where d1 is the minimum distance that UE should be away from the BS.

- Each sector is a cell and that the cell association is geometry based.

- During the simulation, inter-cell handover or switching should be disabled.

For training data generation:

- For each UE moving trajectory: the total length of the UE trajectory can be set as T seconds if it is in time, or set as D meter if it is in distance.

- The trajectory sampling interval granularity depends on UE speed.

- UE can move straight along the entire trajectory, or

- UE can move straight during the time interval, where the time interval is provided by using an exponential distribution with average interval length ΔT

- UE may change the moving direction at the end of the time interval. UE will change the moving direction with the angle difference A\_diff from the beginning of the time interval, provided by using a uniform distribution within [-45°, 45°]

- If the UE trajectory hits the cell boundary (the red line), the trajectory should be terminated.

- If the trajectory length (in time) is less than the length of observation window + prediction window, the trajectory should be discarded.

- The length of observation window + prediction window is not fixed and companies can report their values.

For AI/ML in beam management evaluation, RAN1 does not attempt to define any common AI/ML model as a baseline.

Table 6.3.1-2 presents the baseline link level simulation assumptions for AI/ML in beam management evaluations.

Table 6.3.1-2: Baseline Link Level Simulation assumptions for AI/ML in beam management evaluations

|  |  |
| --- | --- |
| Parameter | Value |
| Frequency | 30GHz. |
| Subcarrier spacing | 120kHz |
| Data allocation | [8 RBs] as baseline, companies can report larger number of RBsFirst 2 OFDM symbols for PDCCH, and following 12 OFDM symbols for data channel |
| PDCCH decoding | Ideal or Non-ideal |
| Channel model | FFS:LOS channel: CDL-D extension, DS = 100nsNLOS channel: CDL-A/B/C extension, DS = 100nsCompanies to explain details of extension methodology considering spatial consistency.Other channel models are not precluded. |
| BS antenna configurations | One panel: (M, N, P, Mg, Ng) = (4, 8, 2, 1, 1), (dV, dH) = (0.5, 0.5) λ as baseline.Other assumptions are not precluded.Companies to explain TXRU weights mapping.Companies to explain beam selection.Companies to explain number of BS beams |
| BS antenna element radiation pattern | Same as SLS |
| BS antenna height and antenna array down-tilt angle | 25m, 110° |
| UE antenna configurations | Panel structure: (M, N, P) = (1, 4, 2), - 2 panels (left, right) with (Mg, Ng) = (1, 2) as baseline- 1 panel as optional- Other assumptions are not precludedCompanies to explain TXRU weights mapping.Companies to explain beam and panel selection.Companies to explain number of UE beams |
| UE antenna element radiation pattern | Same as SLS |
| UE moving speed | Same as SLS |
| Raw data collection format | Depends on sub-use case and companies’ choice.  |

### 6.3.2 Performance results

BM\_Table 1 through BM\_Table 5 in attached Spreadsheets for Beam Management evaluations present the performance results for:

- BM\_Table 1: Evaluation results for BMCase-1 without generalization

- BM\_Table 2: Evaluation results for BMCase-2 without generalization

- BM\_Table 3: Evaluation results for BMCase-1 with generalization for DL Tx beam prediction

- BM\_Table 4. Evaluation results for BMCase-1 with generalization for beam pair prediction

- BM\_Table 5. Evaluation results for BMCase-2 with generalization for DL Tx beam and beam pair prediction

In the evaluation, SLS are used for data generation for training/inference unless otherwise stated.

In the following performance results, Top-K/1(%) is used for Top-K DL Tx beam prediction accuracy or Top-K beam pair prediction accuracy.

Figure 6.3.2-1 and Table 6.3.2-1illustrate model parameter (M) and computational complexity in FLOPs (M) for BM-Case 1 and BM-Case 2, Tx beam prediction and beam pair prediction respectively, according to the reported assumption in BM\_Table 1 and BM\_Table 2.

Note: Optimization of AI/ML model (e.g., in terms of model/computational complexity) was not discussed in the study.

Figure 6.3.2-1: Complexity of AI/ML models from evaluation results
in terms of FLOPs and number of parameters for BM cases

Table 6.3.2-1: AI/ML model complexity/computation complexity
used in the evaluations for AI/ML in beam management

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model complexity in number of model parameters | Model complexity in number of model size | Computational complexity (FLOPs) |
| BM-Case 1 DL Tx beam | More than 1k to 4.9M majority reported less than 1M or about 1M | 50Kbytes to 20Mbytes majority reported less than 0.1Mbytes ~ 0.6Mbytes | ~2.7K to 222M majority reported less than 1M or 10s M  |
| BM-Case 1 DL beam pair | 72k to 4.9M majority reported less than 0.1s M ~ 1M | 0.17Mbytes to 21Mbytes majority reported less than 1Mbytes ~ 4Mbytes | 15K to 224M majority reported less than 1M ~ 4 M |
| BM-Case 2 DL Tx beam | 35k to 11M majority reported less than 0.1s M ~ 1M | 0.5Mbytes to 15Mbytes majority reported about 1s Mbytes  | ~90K to 54M majority reported less than 0.1s M or 1s M |
| BM-Case 2 DL beam pair | 20k to 13Mmajority reported about 0.1M~1M | 0.08M to 15M majority reported about 1Mbytes  | ~90K to 443Mmajority reported less than 0.4 M or 1s M |

In the following performance results, Top-K/1(%) is used for Top-K DL Tx beam prediction accuracy or Top-K beam pair prediction accuracy.

#### 6.3.2.1 Basic performance for BM-Case1

*BM-Case1*: Spatial-domain Downlink beam prediction for Set A of beams based on measurement results of Set B of beams

Note that in the following evaluations, ideal measurements are assumed

- Beams could be measured regardless of their SNR.

- No measurement error.

- Measured in a single-time instance (within a channel-coherence time interval).

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

##### 6.3.2.1.1 Performance when Set B is a subset of Set A for DL Tx beam prediction

For **BM-Case1 DL Tx beam prediction**, when *Set B is a subset of Set A*, AI/ML can provide good beam prediction performance with less measurement/RS overhead comparing to using all measurements of Set A (which provides 100% beam prediction performance as non-AI baseline Option 1) without considering generalization aspects with the measurements from the best Rx beam without UE rotation.

- (A)With measurements of fixed Set B of beams that of 1/4 of Set A of beams

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 9 sources indicate that, AI/ML can achieve about 70%~80% beam prediction accuracy

- evaluation results from 9 sources indicate that, AI/ML can achieve about 80%~90% beam prediction accuracy

- evaluation results from 7 sources indicate that, AI/ML can achieve more than 90% beam prediction accuracy

- evaluation results from 1 source indicates that AI/ML can achieve about 60% beam prediction accuracy when the DL Tx beam grid is generated with oversampling

- Note: 1 source reported that, AI/ML can achieve more than 90% beam prediction accuracy for 100% outdoor UE, and AI/ML can achieve less than 80% beam prediction accuracy for 80% indoor and 20% outdoor. All other results are with the assumption of 80% indoor and 20% outdoor.

- Note: 1 source reported that, AI/ML can achieve 97.3% beam prediction accuracy with the measurements from the best Rx beam based on the best Tx beam in Set A, and AI/ML can achieve 76.4% beam prediction accuracy with the measurements from the best Rx beam of on the best Tx beam in Set B, and 1 source reported that using the best Rx beam in Set A and Set B have similar performance, i.e., 84.84% and 84.59% respectively.

- Non-AI baseline Option 2 (exhaustive beam sweeping in Set B of beams) can achieve about 25% beam prediction accuracy.

- Top-1 DL Tx beam with 1dB margin:

- evaluation results from 15 sources indicate that, AI/ML can achieve more than or about 90% beam prediction accuracy.

- evaluation results from 3 sources indicate that, AI/ML can achieve about 80% beam prediction accuracy, wherein 1 source assumed the L1-RSRP of the Top-1 predicted beam is measured with the best Rx beam searched from the best Tx beam in set B.

- Top-K(=2) DL Tx beam prediction accuracy

- evaluation results from 7 sources indicate that, AI/ML can achieve 80%- 90% beam prediction accuracy.

- evaluation results from 14 sources indicate that, AI/ML can achieve more than 90% beam prediction accuracy.

- The beam prediction accuracy increases with K.

- evaluation results from indicate that Top-2 DL beam prediction accuracy can be more than 95%

- evaluation results from 2 sources indicate that Top-3 DL beam prediction accuracy can be more than 95%

- evaluation results from 3 sources indicate that Top-4 DL beam prediction accuracy can be more than 95%

- evaluation results from 4 sources indicate that Top-5 DL beam prediction accuracy can be more than 95%

- Average L1-RSRP difference of Top-1 predicted beam

- evaluation results from 17 sources indicate that it can be below or about 1dB

- evaluation results from 2 sources indicate that it can be 2.6~2.7dB with the assumption that the L1-RSRP of the Top-1 predicted beam is measured with the best Rx beam searched from the best Tx beam in set B

- Average predicted L1-RSRP difference of Top-1 beam

- evaluation results from 5 sources indicate that it can be below or about 1dB

- evaluation results from 1 source indicates that it is about 2dB

- Note that this is assumed that all the L1-RSRPs of Set A of beams are used as the label in AI/ML training phase (e.g., regression AI/ML model)

- UE average throughput

- evaluation results from 3 sources indicate that AI/ML achieves 96%~99% of the UE average throughput of the BM-Case1 baseline option 1 (exhaustive search over Set A beams).

- evaluation results from 1 source indicates that non-AI baseline option 2 (exhaustive search over Set B beams) achieves 89% of the UE average throughput of the BM-Case1 baseline option 1 (exhaustive search over Set A beams).

- UE 5%ile throughput

- evaluation results from 2 sources indicate that, AI/ML achieves 95~97% of the UE 5%ile throughput of the BM-Case1 baseline option 1 (exhaustive search over Set A beams).

- (B) With measurements of fixed Set B of beams that of 1/8 of Set A of beams

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 7 sources indicate that, AI/ML can achieve about 50% beam prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve about 60%~70% beam prediction accuracy

- evaluation results from 5 sources indicate that, AI/ML can achieve about 70%~80% beam prediction accuracy.

- evaluation results from 4 sources indicate that, AI/ML can achieve more than 80% beam prediction accuracy

- Note: 1 source reported that, AI/ML can achieve 89% beam prediction accuracy with the measurements from the best Rx beam based on the best Tx beam in Set A, and AI/ML can achieve 67.6% beam prediction accuracy with the measurements from the best Rx beam of on the best Tx beam in Set B.

- Non-AI baseline Option 2 (exhaustive beam sweeping in Set B of beams) can achieve about 12.5% beam prediction accuracy

- Top-1 DL Tx beam prediction with 1dB margin

- evaluation results from 7 sources indicate that, AI/ML can achieve 70%-80% beam prediction accuracy

- wherein 1 source assumed the L1-RSRP of the Top-1 predicted beam is measured with the best Rx beam searched from the best Tx beam in set B.

- evaluation results from 1 source indicate that, AI/ML can achieve 80%-90% beam prediction accuracy

- evaluation results from 5 sources indicate that, AI/ML can achieve more than 90% beam prediction accuracy

- Top-K(=2) DL Tx beam prediction accuracy

- evaluation results from 6 sources indicate that, AI/ML can achieve about 70%~ 80% beam prediction accuracy

- evaluation results from 5 sources indicate that, AI/ML can achieve 80%~90% beam prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve 90% beam prediction accuracy for Top-2 DL Tx beam.

- The beam prediction accuracy increases with K.

- evaluation results from 3 sources indicate that Top-3 DL beam prediction accuracy can be more than 95%

- evaluation results from 4 sources indicate that Top-5 DL beam prediction accuracy can be more than 90%

- Average L1-RSRP difference of Top-1 predicted beam

- evaluation results from 8 sources indicate that it can be below or about 1dB

- evaluation results from 4 sources indicate that it can be 1dB~2dB

- evaluation results from 1 source indicates that it can be 3.4dB with the assumption that the L1-RSRP of the Top-1 predicted beam is measured with the best Rx beam searched from the best Tx beam in set B

- Average predicted L1-RSRP difference of Top-1 beam

- evaluation results from 5 sources indicates that it can be 0.8~1.5dB

- Note that 4 sources assumed that all the L1-RSRPs of Set A of beams are used as the label in AI/ML training phase (e.g., regression AI/ML model) and 1 source assumed that only the L1-RSRP of the Top-1 beam in Set A is used as the label in training phase and the result is 0.82 dB.

- UE average throughput

- evaluation results from 1 source indicates that AI/ML achieves 98% of the UE average throughput of the BMCase1 baseline option 1 (exhaustive search over Set A beams).

- evaluation results from 1 source indicates that AI/ML achieves 85% of the UE average throughput of the BMCase1 baseline option 1 (exhaustive search over Set A beams).

- UE 5%ile throughput

- evaluation results from 1 source indicates that, AI/ML achieves 84% of the UE 5%ile throughput of the BMCase1 baseline option (exhaustive search over Set A beams).

- evaluation results from 1 source indicates that, AI/ML achieves 70% of the UE 5%ile throughput of the BMCase1 baseline option (exhaustive search over Set A beams).

##### 6.3.2.1.2 Performance when Set B is different than Set A for DL Tx beam prediction

For **BM-Case1 DL Tx beam prediction**, when *Set B is different than Set A*, with measurements of Set B of wide beams that are 1/4 or 1/6 or 1/8 of Set A beams, AI/ML can provide good beam prediction performance with less measurement/RS overhead comparing to using all measurements of Set A (which provides 100% beam prediction performance as non-AI baseline Option 1) without considering generalization aspects with the measurements from the best Rx beam without UE rotation.

- Top-1 DL Tx beam

- evaluation results from 3 sources indicate that, AI/ML can achieve more than 80% beam prediction accuracy from 5 sources indicate that, AI/ML can achieve more than 55% beam prediction accuracy

- 2 sources reported more than 80% beam prediction accuracy with 100% outdoor UEs, and more than 60% beam prediction accuracy with 20% outdoor UEs.

- Evaluation results from 1 source shows that, with limited measurements (e.g., 1 or 4) of narrow beams in Set A=32, AI/ML can increase 15% or 30% beam prediction accuracy [respectively] compared with 55% beam prediction accuracy with measurement of wide beams only.

- Top-1 DL Tx beam with 1dB margin

- evaluation results from 4 sources indicate that, AI/ML can achieve more than 85% beam prediction accuracy

- evaluation results from 3 sources indicate that, AI/ML can achieve 57%~77% beam prediction accuracy

- One source reported more than 86% beam prediction accuracy with 100% outdoor UEs, and more than 70% beam prediction accuracy with 20% outdoor UEs.

- Top-K(=3) DL Tx beam

- evaluation results from 3 sources indicate that, AI/ML can achieve more than 95% beam prediction accuracy

- evaluation results from 3 sources indicate that, AI/ML can achieve 85~94% beam prediction accuracy

- evaluation results from 1 source indicates that Top-5 DL beam prediction accuracy can be more than 90%.

- Average L1-RSRP difference of Top-1 predicted beam

- evaluation results from 4 sources indicate that, the average L1-RSRP difference can be less or about 1dB

- UE average throughput

- evaluation results from 1 source indicates that, AI/ML achieves 99% of the UE average throughput of the BMCase1 baseline option 1 (exhaustive search over Set A beams)

- UE 5%ile throughput

- evaluation results from 1 source indicates that, AI/ML achieves 94% of the of the BMCase1 baseline option 1(exhaustive search over Set A beams)

##### 6.3.2.1.3 Performance when Set B is a subset of Set A for DL Tx-Rx beam pair prediction

For **BM-Case1 DL Tx-Rx beam pair prediction**, when *Set B is a subset of Set A*, AI/ML can provide good beam prediction performance with less measurement/RS overhead comparing to using all measurements of Set A (which provides 100% beam prediction performance as non-AI baseline Option 1) without considering generalization aspects and without UE rotation.

- (A) With measurements of fixed Set B of beam pairs that of 1/4 of Set A of beam pairs

- Top-1 beam pair prediction accuracy:

- evaluation results from 8 sources indicate that, AI/ML can achieve about 50%~70% prediction accuracy

- evaluation results from 4 source indicate that, AI/ML can achieve 70%~80% prediction accuracy

- evaluation results from 5 sources indicate that, AI/ML can achieve about 80%~90% prediction accuracy

- evaluation results from 1 source indicates that, AI/ML can achieve more than 90% prediction accuracy

- Note: in the above evaluation and the rest of other KPIs, most of the sources used measurements from all Rx beams of a certain set of Tx beams, except 3 sources who use measurements from half of Rx beams of a certain set of Tx beams.

- The results from 3 sources indicate 60%~68% prediction accuracy in terms of Top-1 beam pair prediction accuracy.

- 1 source additionally reports that, AI/ML can achieve 76.46% and 56.12% beam prediction accuracy with the measurements from all Rx beams and half of Rx beams of a certain set of Tx beams respectively.

- Non-AI baseline Option 2 (exhaustive beam sweeping in Set B of beam pairs) can achieve about 25% prediction accuracy.

- Top-1 beam pair prediction accuracy with 1dB margin:

- evaluation results from 5 sources indicate that, AI/ML can achieve more than 70% prediction accuracy

- evaluation results from 2 sources indicate that, AI/ML can achieve 80%~ about 90% prediction accuracy

- evaluation results from 6 sources indicate that, AI/ML can achieve more than 90% prediction accuracy.

- Note: 1 source reported that, AI/ML can achieve 91.6% and 74.57% beam prediction accuracy with 1dB margin with the measurements from all Rx beams of a certain set of Tx beams and with half of Rx beams of a certain set of Tx beams respectively.

- Top-K(=2) beam pair prediction accuracy

- evaluation results from 2 sources indicate that, AI/ML can achieve 65%- 75% prediction accuracy.

- evaluation results from 6 sources indicate that, AI/ML can achieve 80%- 90% prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve more than 90% prediction accuracy

- Note: 1 source reported that, AI/ML can achieve 91.34% and 78.06% Top-K(=2) beam prediction accuracy with the measurements from all Rx beams and half of Rx beams of a certain set of Tx beams respectively.

- The beam prediction accuracy increases with K.

- evaluation results from 1 source indicate that Top-3 beam pair prediction accuracy can be more than 95%

- evaluation results from 4 sources indicate that Top-4 beam pair prediction accuracy can be [more than 95%

- evaluation results from 2 sources indicate that Top-5 beam pair prediction accuracy can be more than 95%

- evaluation results from 1 source indicate that Top-10 beam pair prediction accuracy can be more than 95% for 32 Tx and 4 Rx with results from half Rx

- Average L1-RSRP difference of Top-1 predicted beam pair

- evaluation results from 13 sources indicate that it can be below or about 1dB

- evaluation results from 1 source indicate that it can be about 1.5dB

- Note: 1 source reported that it can be 0.716dB and 1.611dB with the measurements from all Rx beams and half of Rx beams of a certain set of Tx beams respectively.

- Predicted L1-RSRP difference of Top-1 beam pair

- 3 sources indicate that it can be below or about 1dB

- Note that this is assumed that all the L1-RSRPs of Set A of beams are used as the label in AI/ML training phase (e.g., regression AI/ML model)

- (B) With measurements of fixed Set B of beam pairs that of 1/8 of Set A of beam pairs

- Top-1 beam pair prediction accuracy:

- evaluation results from 4 sources indicate that, AI/ML can achieve about 50% prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve about 60%~70% prediction accuracy

- evaluation results from 6 sources indicate that, AI/ML can achieve about 70%~80% prediction accuracy

- Note: in the above evaluation and the rest of other KPIs, most of the sources used measurements from all Rx beams of a certain set of Tx beams, except 7 sources who use measurements from half of Rx beams of a certain set of Tx beams.

- Non-AI baseline Option 2 (exhaustive beam sweeping in Set B of beam pairs) can achieve about 12.5% prediction accuracy

- Top-1 beam pair prediction with 1dB margin

- evaluation results from 4 sources indicate that, AI/ML can achieve 60%-70% prediction accuracy

- evaluation results from 1 source indicate that, AI/ML can achieve 70%-80% prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve 80%-90% prediction accuracy

- Top-K(=2) beam pair prediction accuracy

- evaluation results from 4 sources indicate that, AI/ML can achieve about 70%- 80% prediction accuracy.

- evaluation results from 6 sources indicate that, AI/ML can achieve 80%- 90% prediction accuracy

- evaluation results from 2 sources indicate that, AI/ML can achieve more than 90% prediction accuracy

- The beam prediction accuracy increases with K.

- evaluation results from 1 source indicate that Top-3 beam pair prediction accuracy can be 96%

- evaluation results from 1 source indicate that Top-4 beam pair prediction accuracy can be 96%

- evaluation results from 1 source indicate that Top-5 beam pair prediction accuracy can be 91%

- evaluation results from 1 source indicate that Top-5 beam pair prediction accuracy can be 94%

- Average L1-RSRP difference of Top-1 predicted beam pair

- evaluation results from 5 sources indicate that it can be below or about 1dB

- evaluation results from 5 sources indicate that it can be 1dB~2dB

- Average predicted L1-RSRP difference of Top-1 beam pair

- evaluation results from 2 sources indicate that it can be 0.7~1.3dB

- Note that this is assumed that all the L1-RSRPs of Set A of beams are used as the label in AI/ML training phase (e.g., regression AI/ML model).

- (C) With measurements of fixed Set B of beams that of 1/16 of Set A of beams

- Top-1 beam pair prediction accuracy

- evaluation results from 5 sources indicate that, AI/ML can achieve less than 50% or about 50% prediction accuracy

- evaluation results from 2 source indicate that, AI/ML can achieve about 55%~57% prediction accuracy

- evaluation results from 3 sources indicate that, AI/ML can achieve about 60%~70% prediction accuracy

- evaluation results from 1 source indicate that, AI/ML can achieve about 70%~80% prediction accuracy

- Note: in the above evaluation and the rest of other KPIs, some 6 sources used measurements from all Rx beams of a certain set of Tx beams, and some other 6 sources use measurements from half or fourth of Rx beams of a certain set of Tx beams.

- Non-AI baseline Option 2 (exhaustive beam sweeping in Set B of beam pairs) can achieve about 6.25% prediction accuracy

- Top-1 beam pair prediction with 1dB margin

- evaluation results from 4 sources indicate that, AI/ML can achieve less than 50% or about 50% prediction accuracy

- evaluation results from 1 source indicate that, AI/ML can achieve more than 50%~60% prediction accuracy

- evaluation results from 3 sources indicate that, AI/ML can achieve about 60%-70% prediction accuracy

- evaluation results from 2 sources indicate that, AI/ML can achieve 72%~85% prediction accuracy

- Top-K(=2) beam pair prediction accuracy

- evaluation results from 3 sources indicate that, AI/ML can achieve less than 60% prediction accuracy.

- evaluation results from 5 sources indicate that, AI/ML can achieve about 70%- 80% prediction accuracy

- evaluation results from 1 source indicate that, AI/ML can achieve more than 85% prediction accuracy

- The beam prediction accuracy increases with K.

- Average L1-RSRP difference of Top-1 predicted beam pair

- evaluation results from 3 sources indicate that it can be 1dB~2dB

- evaluation results from 2 sources indicate that it can be 2dB~3dB

- evaluation results from 2 sources indicate that it can be more than 3dB

- evaluation results from 1 source indicate that it can be about 6dB

- Predicted L1-RSRP difference of Top-1 beam pair

- evaluation results from 2 sources indicates that it can be about 2.5dB

- Note that this is assumed that all the L1-RSRPs of Set A of beams are used as the label in AI/ML training phase (e.g., regression AI/ML model).

- Note: in the above evaluations, 8 sources assumed 4 Rx, other sources assumed 8 Rx.

##### 6.3.2.1.4 Performance when Set B is different to Set A for DL Tx-Rx beam pair prediction

**For BM-Case1 beam pair prediction**, when *Set B is different to Set A*, with measurements of Set B of Tx wide beams that are 1/4 or 1/8 of Set A beams, evaluation results from 1 source indicate that AI/ML can provide good beam prediction performance with less measurement/RS overhead compared to using all measurements of Set A (which provides 100% beam prediction performance as non-AI baseline Option 1) without considering generalization and without UE rotation.

- For Top-1 beam pair prediction accuracy, evaluation results from 1 source indicate that, AI/ML can achieve about 92.7%/92.5% beam prediction accuracy for 1/4 and 1/8 overhead respectively.

- For Top-1 beam prediction accuracy with 1dB margin, evaluation results from 1 source indicate that, AI/ML can achieve about 97.6%/97.3% beam prediction accuracy for 1/4 and 1/8 overhead respectively.

#### 6.3.2.2 Basic performance for BM-Case2

*BM-Case2:* Temporal Downlink beam prediction for Set A of beams based on the historic measurement results of Set B of beams.

##### 6.3.2.2.1 Performance when Set A = Set B

**For BM-Case2**, when *Set B = Set A*, for DL Tx beam prediction with the measurements from the best Rx beam or Tx-Rx beam pair prediction, without considering generalization aspects, with the following assumptions:

- UE speed: 30km/h (unless otherwise stated)

- Prediction time: 80ms/160ms/320ms/640ms/800ms/others

- With UE rotation and without UE rotation

- Set B is the same as Set A in each time instance for measurement

Note that ideal measurements are assumed.

- Beams could be measured regardless of their SNR.

- No measurement error.

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

**(A) For Tx DL beam prediction,** based on most of the evaluation results, AI/ML provides some beam prediction accuracy gain for prediction time larger than or equal to 160ms, and some evaluation results show AI/ML may have similar performance or some degradation for 80ms or 160ms prediction time comparing with non-AI baseline (Option 2, sample and hold based on the previous measurements) with same RS/measurement overhead **without UE rotation**. For the longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML:

- For 80ms prediction time, evaluation results from 1 source show that AI/ML may **have similar performance or may decrease** about 4% beam prediction accuracy, evaluation results from 2 sources show that AI/ML may have similar performance or may **decrease** 0.4%~1% beam prediction accuracy, evaluation results from 1 source show that AI/ML can **increase** about 1%~2% prediction accuracy in terms of Top-1 beam prediction accuracy,

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 40ms. And it can decrease 4% beam prediction accuracy comparing with 98.23% achieved by non-AI baseline (Option 2-2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms/160ms. And it may decrease up to 0.4~1% beam prediction accuracy comparing with about 80%/78.7% achieved by non-AI baseline (Option 2) with 32 Tx beams.

- wherein, 1 source used measurements from 8 time instances with measurement periodicity of 40ms. And it can decrease about 0.5% beam prediction accuracy comparing with 67.4% achieved by non-AI baseline (Option 2) with 64 Tx beams

- wherein, 1 source used measurements from 5 time instances with measurement periodicity of 80ms. And it can increase 1% beam prediction accuracy gain comparing with 78.5% and 76.2% achieved by non-AI baseline (Option 2) with 32 Tx beams for 30km/h and 60km/h respectively.

- For 160ms prediction time, evaluation results from 3 sources show that AI/ML may have similar performance or may decrease 1%~5% beam prediction accuracy in terms of Top-1 beam prediction accuracy, evaluation results from 3 sources show that AI/ML can increase 1%~2% prediction accuracy, evaluation results from 3 sources show that AI/ML can increase 4%~5% prediction accuracy and evaluation results from 2 sources show that AI/ML can increase about 10% prediction accuracy in terms of Top-1 beam prediction accuracy.

- wherein, 1 source used measurements from 3 time instances with measurement periodicity of 80ms. And AI/ML does not provide beam prediction accuracy gain comparing with 83.9% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 40ms. And it can decrease 5% beam prediction accuracy comparing with 97.18% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms/160ms/240ms/320ms. And it may decrease up to 2% beam prediction accuracy comparing with about 73.8%~80.9%% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 6 time instances with measurement periodicity of 40ms. And it can increase 4% beam prediction accuracy comparing with achieved 64.4% by non-AI baseline (Option 2) with 60km/h UE speed and 32 Tx beams

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 160ms. And it can increase 4% beam prediction accuracy comparing with 52% achieved by non-AI baseline (Option 2) with 64 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 160ms. And it can increase 5% beam prediction accuracy comparing with 61.2% achieved by non-AI baseline (baseline 2) with 32 Tx beams

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 80ms. And it can increase 1.9% beam prediction accuracy comparing with 93.2% achieved by non-AI baseline (baseline 2) with 32 Tx beams

- wherein, 1 source used measurements from 5 time instances with measurement periodicity of 160ms. And it can increase 10.8% beam prediction accuracy comparing with achieved 82.2% by non-AI baseline (Option 2) with 30km/h UE speed and 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 40ms. And it can increase 1% beam prediction accuracy comparing with 85.8% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 8 time instances with measurement periodicity of 40ms. And it can increase about 2% beam prediction accuracy comparing with 67.4% achieved by non-AI baseline (Option 2) with 64 Tx beams

- wherein, 1 source used measurements from 3 time instances with measurement periodicity of 160ms. And it can increase about 9.2% and about 4.6% beam prediction accuracy comparing with 51.36% and 45.76% achieved by non-AI baseline (Option 2) with 30km/h and 60km/h UE speed respectively with 64 Tx beams

- For 320ms prediction time, evaluation results from 7 sources show that AI/ML can increase about up to 3%~8% prediction accuracy, and evaluation results from 2 sources show that AI/ML can increase about 18.5%~23.5% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 160ms. And it can increase 6% beam prediction accuracy comparing with 39.7% achieved by non-AI baseline (Option 2) with 64 Tx beams.

- wherein, 1 source used measurements from 6 time instances with measurement periodicity of 80ms. And it can increase 8% beam prediction accuracy comparing with achieved 55.5% by non-AI baseline (Option 2) with 60km/h UE speed and for 32 Tx beams

- wherein, 1 source used measurements from 3 time instances with measurement periodicity of 160ms. And it can increase 18.5% and 23.5% beam prediction accuracy comparing with 42.78% and 34.53% achieved by non-AI baseline (Option 2) with 30km/h and 60km/h UE speed respectively and for 64 Tx beams.

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 320ms. And it can increase 3.5% beam prediction accuracy comparing with 60.82% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 80ms. And it can increase 3.2% beam prediction accuracy comparing with 90.1% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 5 time instances with measurement periodicity of 160ms. And it can increase 18.4% beam prediction accuracy comparing with 74.4% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms. And it can increase 4.2% beam prediction accuracy comparing with 79.4% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms/160ms/320ms/400ms /480ms/640ms. And it can increase up to 3.4% beam prediction accuracy comparing with about 69.5~78.5% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 8 time instances with measurement periodicity of 40ms. And it can increase about 3% beam prediction accuracy comparing with 29.1% achieved by non-AI baseline (Option 2) with 64 Tx beams

- For 640ms prediction time, evaluation results from 5 sources show that AI/ML can increase 4.5~8% prediction accuracy, and evaluation results from 1 source show that AI/ML can increase up to 14.3% prediction accuracy in terms of Top-1 beam prediction accuracy, and evaluation results from 1 source show that AI/ML can increase up to 28.5% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 160ms. And it can increase 8% beam prediction accuracy comparing with 35.2% achieved by non-AI baseline (Option 2) with 64 Tx beams

- wherein, 1 source used measurements from 6 time instances with measurement periodicity of 160ms. And it can increase 14.3% beam prediction accuracy comparing with achieved 41.8% by non-AI baseline (Option 2) with 60km/h UE speed and 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 320ms. And it can increase 4.5% beam prediction accuracy comparing with 58% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 2 time instances with measurement periodicity of 80ms. And it can increase 5.4% beam prediction accuracy comparing with 84.4% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 5 time instances with measurement periodicity of 160ms. And it can increase 28.5% beam prediction accuracy comparing with 63.9% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 160ms. And it can increase 7.8% beam prediction accuracy comparing with 67.9% achieved by non-AI baseline (Option 2) with 32 Tx beams

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 160ms/320ms/640ms/800ms/960ms/1280ms. And it can increase up to 8.2% beam prediction accuracy comparing with about 62.7~74.3% achieved by non-AI baseline (Option 2) with 32 Tx beams

- For 800ms prediction time, in terms of Top-1 beam prediction accuracy

- evaluation results from 1 source show that AI/ML can increase about 3.5% prediction accuracy comparing with 34.6% achieved by non-AI baseline (Option 2) with 64 Tx beams with measurements from 2 time instances in measurement periodicity of 160ms

- evaluation results from 1 source show that AI/ML can increase about 33.7% prediction accuracy comparing with achieved 58.6% by non-AI baseline (Option 2) 32 Tx beams with measurements from 5 time instances with measurement periodicity of 160ms

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 800ms/1600ms. And it can increase up to 9.1% beam prediction accuracy comparing with about 61.5~66.5% achieved by non-AI baseline (Option 2) with 32 Tx beams

- For 960ms prediction time, in terms of Top-1 beam prediction accuracy

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 960ms/1920ms. And it can increase up to 10.6% beam prediction accuracy comparing with about 60.1~64.4% achieved by non-AI baseline (Option 2) with 32 Tx beams

- For 1280ms prediction time, in terms of Top-1 beam prediction accuracy

- evaluation results from 1 source show that AI/ML can increase about 12.7% beam prediction accuracy comparing with 54.3% achieved by non-AI baseline (Option 2) with 32 Tx beams with measurements from 4 time instances with measurement periodicity of 320ms.

- evaluation results from 1 source show that AI/ML can increase about 4%~13.4% beam prediction accuracy comparing with 54%~66.8% achieved by non-AI baseline (Option 2) with 32 Tx beams with measurements from 4 time instances with measurement periodicity of 320ms~2560ms.

- evaluation results from 1 source show that AI/ML can increase up to 17.6% prediction accuracy for 3200ms prediction time.

- evaluation results from 1 source show that AI/ML can increase up to 19.1% prediction accuracy for up to 12.8s prediction time.

- Beam prediction accuracy gain in terms of Top-K prediction accuracy or Top-1 prediction accuracy with 1dB error is similar as or smaller than the beam prediction accuracy gain in terms of Top-1 prediction accuracy.

- For the prediction time no larger than 1280ms, AI/ML and non-AI baseline (Option 2) can provide similar average L1-RSRP error, which are less than 1dB.

**(B) For Tx DL beam prediction,** based on the evaluation from 2 sources, AI/ML **can** provide some beam prediction accuracy gain comparing with non-AI baseline (Option 2, sample-and-hold) **with UE rotation** andthe performance of AI/ML compared to baseline (Option 2, sample-and-hold) improves with the increase of measurement periodicity:

**- For 160ms/800ms/1200ms/1600ms prediction time,** evaluation results from 1 source show about 2%/8%/10%/13% prediction accuracy increase comparing with 74%/60%/53%/47.7% achieved by non-AI baseline (Option 2) with 32 Tx beam respectively in terms of Top-1 beam prediction accuracy, with measurements from 4 time instances in measurement periodicity of 160ms/800ms/ 1200ms/1600ms respectively.

- In the evaluation, UE rotation is modelled every 20ms with a rotation speed uniformly distributed within {0, 60} RPM, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

**- For 160ms/320ms/480ms/960ms prediction time,** evaluation results from 1 source show that AI/ML can increase 2%/3%/4.2%/7.3% Top-1 beam prediction accuracy compared to non-AI baseline (Option 2) with 78%/75.5%/73%/66.3% beam prediction accuracy with 12 Tx with measurement periodicity of 200ms/360ms/520ms/1000ms.

- In the evaluation, UE rotation is modelled every 40ms with constant 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes.

**(C) For Tx DL beam prediction** **(without UE rotation unless otherwise stated**), AI/ML can provide good beam prediction accuracy with the less measurements/RS overhead:

- Under the assumption of **setting Case A,** decent beam prediction accuracy can be achieved with **1/5~1/2** measurement/RS overhead reduction comparing the non-AI baseline (Option 1, with 100% prediction accuracy)

- evaluation results from 1 source show that AI/ML can achieve 57% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 52% beam prediction accuracy in term of Top-1 beam prediction accuracy for 160ms prediction time,

**- 1/3 RS/measurement overhead reduction** can be obtained with measurements from 2 time instances with measurement periodicity of 160ms.

- When prediction time increased to 320ms or larger, >50% Top-1 beam prediction accuracy is lower than 50% even with the help of AI/ML although it still can provide some gain compared with non-AI baseline (Option2).

- evaluation results from 1 source show that AI/ML can achieve 60%~71% beam prediction accuracyin terms of Top-1 beam prediction accuracy for 40ms up to 240ms prediction time

**- 3/7 RS/measurement overhead reduction** can be obtained with measurements from 8 time instances with measurement periodicity of 40ms.

- When prediction time increased to 280ms or larger, >50% Top-1 beam prediction accuracy is lower than 50% even with the help of AI/ML

- evaluation results from 1 source show that AI/ML can achieve 60.5% beam prediction accuracyin terms of Top-1 beam prediction accuracy for up to 320ms prediction time

**- 2/5 RS/measurement overhead reduction** can be obtained with measurements from 3 time instances with measurement periodicity of 160ms.

- evaluation results from 1 source show that AI/ML can achieve 86.8%/83.6%/75.7%/67% beam prediction accuracyin terms of Top-1 beam prediction accuracy for up to 160ms/320ms/640ms/1280ms prediction time, respectively

**- 1/2 RS/measurement overhead reduction** can be obtained with measurements from 4 time instances with measurement periodicity of 40ms/80ms/160ms/320ms, respectively.

- evaluation results from 1 source show that AI/ML can achieve 92% beam prediction accuracy in terms of Top-1 beam prediction accuracy for 160ms up to 800ms prediction time

**- 1/2 RS/measurement overhead reduction** can be obtained with measurements from 5 time instances with measurement periodicity of 160ms.

- evaluation results from 1 source show that AI/ML can achieve 64%~68%/56%~63%/ 47%~56% beam prediction accuracyin terms of Top-1 beam prediction accuracy for 160ms/320ms/ 640ms prediction time respectively

**- 2/5 RS/measurement overhead reduction** can be obtained with measurements from 5 time instances with measurement periodicity of 40ms/80ms/160ms respectively.

- evaluation results from 1 source show that AI/ML can achieve 62%~66% beam prediction accuracyin terms of Top-1 beam prediction accuracy for 160ms to 640ms prediction time

**- 1/5 RS/measurement overhead reduction** can be obtained with measurements from 4 time instances with measurement periodicity of 160ms to 640ms.

- evaluation results from 1 source show that AI/ML can achieve 58.0%~80.1% beam prediction accuracyin terms of Top-1 beam prediction accuracy for 160ms to 12800ms prediction time

**- up to 1/2 RS/measurement overhead reduction** can be obtained with measurements from 4 time instances with measurement periodicity of 160ms to 3200ms.

- Under the assumption of **setting Case B,** evaluation results from 2 sources indicate that a certain beam prediction accuracy can be achieved with 1/2 ~ 7/10 measurement/RS overhead reduction comparing with non-AI schemes (Option 2)

- evaluation results from 1 source show that AI/ML can provide **1/2 RS/measurement overhead reduction with UE rotation:**

- AI/ML can achieve ~65% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 48% beam prediction accuracy in term of Top-1 beam prediction accuracy for 1600ms prediction time/measurement periodicity

- With non-AI baseline (Option 2), similar prediction accuracy (~65% of Top-1 beam prediction accuracy) can be achieved with 800ms prediction time /measurement periodicity.

- In the evaluation, **UE rotation** is modelled every 20ms with a rotation speed of RPM = 60 R/M, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

- evaluation results from 1 source show that AI/ML can provide **7/10 RS/measurement overhead reduction without UE rotation:**

- AI/ML can achieve ~64% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 46% beam prediction accuracy in term of Top-1 beam prediction accuracy for 3200ms prediction time

- With non-AI baseline (Option 2), similar prediction accuracy (~64% of Top-1 beam prediction accuracy) can be achieved with 960ms prediction time.

- Under the assumption of **setting Case B+,** based on the evaluation results from 2 sources, good beam prediction accuracy can be achieved by AI/ML with measurement/RS overhead reduction compared to the non-AI baseline (Option 1, with 100% prediction accuracy) for which minimal periodicity of measurement is Tper

- evaluation results from 1 source with Tper = 40ms show that AI/ML can provide 80%/88.9%/92.3%/96% RS/measurement overhead reduction:

- AI/ML can achieve 80%/78.5%/77.2%/73.6% beam prediction accuracy in terms of Top-1 beam prediction accuracy with 160ms/320ms/480ms/960ms prediction time 200ms/360ms/520ms/ 1000ms measurement periodicity.

- In the evaluation, UE rotation is modelled every 40ms with constant 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes.

- evaluation results from 1 source with Tper = 160ms~3200ms show that AI/ML can provide 80% RS/measurement overhead reduction:

- AI/ML can achieve 50%~73% beam prediction accuracy in terms of Top-1 beam prediction accuracy with 640ms to 12800ms prediction time (4 prediction time instance) /800ms to 16000ms measurement periodicity (4 measurement time instance) without UE rotation.

**(D) For beam pair prediction,** AI/ML may or may not provide beam prediction accuracy gain comparing with non-AI baseline (Option 2) for 160ms or less prediction time **without UE rotation.** For the longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML:

- For 160ms prediction time, evaluation results from 2 sources show AI/ML can provide similar performance or increase up to 1% prediction accuracy gain, evaluation results from 1 source show AI/ML may decrease 8% prediction accuracy, and evaluation results from 1 source show AI/ML can increase 13.8% prediction accuracy, in terms of Top-1 beam prediction accuracy.

- evaluation results from 1 source show that AI/ML decrease 8% prediction accuracy in terms of Top-1 beam prediction accuracy with measurements from 4 time instances with measurement periodicity of 160ms comparing with 68.1% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- evaluation results from 1 source show that AI/ML can increase 0.1% beam prediction accuracy in terms of Top-1 beam prediction accuracy with measurements from 4 time instances with measurement periodicity of 40ms comparing with 81.3% achieved by non-AI baseline (Option 2) with 32 Tx beams and 4 Rx beams.

- evaluation results from 1 source show that AI/ML decrease 0.1%~1% prediction accuracy in terms of Top-1 beam prediction accuracy with measurements from 4 time instances with measurement periodicity of 80ms~320ms comparing with 80.7%~83.4% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- evaluation results from 1 source show that AI/ML can increase 13.8% prediction accuracy in terms of Top-1 beam prediction accuracy with measurements from 5 time instances with measurement periodicity of 160ms comparing with 78.1% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- For 320ms prediction time, evaluation results from 2 sources show that AI/ML can increase less than 3% prediction accuracy in terms of Top-1 beam prediction accuracy, and evaluation results from 1 source show that AI/ML can increase 22.5% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms~640ms. With one AI/ML model to predict the beam at one or multiple time instances including 320ms, AI/ML may increase [less than 2%] beam prediction accuracy comparing with 78.8%~81.2% achieved by non-AI baseline (Option 2)

- Wherein, 1 source used measurements from 4 time instances with measurement periodicity of 80ms and it shows that AI/ML can increase 2.8% beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing with 74.5% achieved by non-AI baseline (Option 2).

- Wherein, 1 source used measurements from 5 time instances with measurement periodicity of 160ms and it shows that AI/ML can increase 22.5% prediction accuracy in terms of Top-1 beam prediction accuracy comparing with 69.2% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- For 640ms prediction time, evaluation results from 2 sources show that AI/ML may be able to increase up to 7.5% prediction accuracy, and evaluation results from 1 source show that AI/ML can increase 34% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, 1 source used measurements from 4 time instances

- With one AI/ML model to predict the beam at 640ms with 640/1280ms as measurement periodicity, AI/ML can increase 6%/3.5% beam prediction accuracy comparing with 74.1%/73.5% achieved by non-AI baseline (Option 2).

- With one AI/ML model to predict the beam at multiple prediction time instances (with two or more of 160ms 320ms, 480ms, 640ms) with different measurement periodicities (e.g., 160ms, 320ms, 800ms, 960ms), AI/ML can increase [0.7%~3.5%] beam prediction accuracy. From the evaluation results, the more target predicted time instances, the less performance gain can be obtained from AI/ML.

- Wherein, 1 source used measurements from 4 time instances with measurement periodicity of 160ms and it shows that AI/ML can increase 7.5% beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing with 63.3% achieved by non-AI baseline (Option 2)

- Wherein, 1 source used measurements from 5 time instances with measurement periodicity of 160ms and it shows that AI/ML can increase 34% prediction accuracy in terms of Top-1 beam prediction accuracy comparing with 57.16% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- For 800ms prediction time,

- evaluation results from 1 source show that AI/ML can to increase 6.7%~7.5% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, measurements from 4 time instances with 800ms/1600ms as measurement periodicity were used and AI/ML can increase 6.7%/7.5% beam prediction accuracy respectively comparing with 72.9%/69.2%achieved by non-AI baseline (Option 2).

- evaluation results from 1 source show that AI/ML can to increase 39.4% prediction accuracy in terms of Top-1 beam prediction accuracy

- wherein, measurements from 5 time instances with 160ms as measurement periodicity were used and AI/ML can increase 39.4% beam prediction accuracy comparing with 51.2% achieved by non-AI baseline (Option 2) with 32 Tx beams and 8 Rx beams.

- For 960ms prediction time,

- evaluation results from 1 source show that AI/ML may increase 12.8% beam prediction accuracy in terms of Top-1 beam prediction accuracy

- Wherein measurements from 5 time instances with measurement periodicity of 160ms, and predictions of 95 time instances with prediction periodicity of 10ms are assumed. AI/ML has 12.8% of beam prediction accuracy improvement in terms of Top 1 beam prediction accuracy comparing with 57.5% achieved by non-AI baseline (Option 2).

- evaluation results from 1 source show that AI/ML may be able to increase up to 8.5% prediction accuracy in terms of Top-1 beam prediction accuracy

- measurements from 4 time instances with measurement periodicity of 960ms/1920ms were used respectively, with one model to predict single /multiple prediction time instances. AI/ML can increase 8.1%/8.5% beam prediction accuracy respectively comparing with 71.3%/67.7%achieved by non-AI baseline (Option 2).

- For 1200ms/1600ms/2400ms/3200ms/40000ms prediction time, evaluation results from 1 source show that AI/ML may be able to increase up to 8.8%/ up to 10.7%/ up to 10.2%/up to 11.3%/up to 20.4% prediction accuracy in terms of Top-1 beam prediction accuracy respectively

- measurements from 4 time instances were used with 1200ms/1600ms /1200ms/1600ms/4000ms as measurement periodicity respectively

**(E)For beam pair prediction**, based on the evaluation results from 3 sources, AI/ML **may or may not** provide beam prediction accuracy gain comparing with non-AI baseline (Option 2) **with UE rotation:**

- For 160ms prediction time, in terms of Top-1 beam prediction accuracy

- evaluation results from 1 source show that AI/ML may decrease 10% prediction accuracy with measurements from 4 time instances with measurement periodicity of 160ms. In this case, non-AI baseline (option 2) can achieve 51.09% beam prediction accuracy.

- In the evaluation, UE rotation is modelled every 20ms with a rotation speed uniformly distributed within {0, 60} RPM, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

- For 200ms prediction time, in terms of Top-1 beam prediction accuracy with 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes

- evaluation results from 1 source show that AI/ML can increase [1%~1.6%] prediction accuracy with measurement periodicity of 240ms with different AI/ML models. In this case, non-AI baseline (option 2) can achieve 67.4% beam prediction accuracy

- For 200ms prediction time, in terms of Top-1 beam prediction accuracy with 100 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes

- evaluation results from 1 source show that AI/ML can increase 23%~30% prediction accuracy with measurement periodicity of 240ms with different AI/ML models. In this case, non-AI baseline (option 2) can only achieve 17% beam prediction accuracy.

- For 500ms prediction time, in terms of Top-1 beam prediction accuracy with 10 RPM rotation speed to fixed a direction

- evaluation results from 1 source show that AI/ML can increase 6%/8%/11% prediction accuracy with measurements from 1/2/5 time instances in measurement periodicity of 100ms respectively

- evaluation results from 1 source show that AI/ML can increase 11%/11.5%/12.5% prediction accuracy with measurements from 1/2/5 time instances in measurement periodicity of 50ms respectively

- For 800ms prediction time, in terms of Top-1 beam prediction accuracy

- evaluation results from 1 source show that AI/ML may decrease 6% prediction accuracy with measurements from 4 time instances with measurement periodicity of 800ms. In this case, non-AI baseline (option 2) can achieve 30.19% prediction accuracy.

- In the evaluation, UE rotation is modelled every 20ms with a rotation speed uniformly distributed within {0, 60} RPM, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

**(F) For beam pair prediction,** (without UE rotation unless otherwise stated), AI/ML can provide good beam prediction accuracy with the less measurements/RS overhead:

- Under assumption of **setting Case A,** decent beam prediction accuracy can be achieved with up to 1/2 measurement/RS overhead comparing with no time domain prediction.

- evaluation results from 1 source show that AI/ML can achieve 81.4%/77.3%/70.8%/61.8% beam prediction accuracyin terms of Top-1 beam prediction accuracy for up to 160ms/320ms/640ms/1280ms prediction time, respectively

**- 1/2 RS/measurement overhead reduction** can be obtained with measurements from 4 time instances with measurement periodicity of 40ms/80ms/160ms/320ms.

- evaluation results from 1 source show that AI/ML can achieve 90%-92% beam prediction accuracy in terms of Top-1 beam prediction accuracy for 160ms up to 800ms prediction time

**- 1/2 RS/measurement overhead reduction** can be obtained with measurements from 5 time instances with measurement periodicity of 160ms.

- evaluation results from 1 source show that AI/ML can achieve 79%~84% beam prediction accuracyin terms of Top-1 beam prediction accuracy for 80ms to 640ms prediction time without UE rotation for beam pair

**- up to 1/2 RS/measurement overhead reduction** can be obtained with measurements from 4 time instances with measurement periodicity of 80ms or 160ms.

- evaluation results from 1 source show that AI/ML can achieve 71.9% /67.4%/64.4% for 30km/h /60km/h /90km/h beam prediction accuracyrespectivelyin terms of Top-1 beam prediction accuracy for 800ms prediction time.

**- 1/2** RS/measurement overhead reduction can be obtained with measurements from 5 time instances with measurement periodicity of 160ms.

- Under assumption of **setting Case B**, based on the evaluation from 2 sources a certain beam prediction accuracy can be achieved with 1/2 or 3/5 measurement/RS overhead reduction comparing with non-AI schemes with 30km/h respectively

- evaluation results from 1 source show that AI/ML can provide 1/2 or 2/3 or 3/4 RS/measurement overhead reduction without UE rotation for 30km/h /60km/h /90km/h respectively

- AI/ML can achieve 70.3%/77.1%/79.8% beam prediction accuracy with 30km/h /60km/h /90km/h respectively, while non-AI baseline (Option 2) can only achieve 57.2%/36%/36% beam prediction accuracy in term of Top-1 beam prediction accuracy for 960ms/960ms/640ms prediction time/measurement periodicity for 30km/h /60km/h /90km/h respectively.

- With non-AI baseline (Option 2), similar prediction accuracy (76.7% of Top-1 beam prediction accuracy) can be achieved with 480ms/320ms/160ms measurement periodicity for 30km/h /60km/h /90km/h respectively.

- evaluation results from 1 source show that AI/ML can provide 3/5 RS/measurement overhead reduction without UE rotation

- AI/ML can achieve 77.6% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 66.9% beam prediction accuracy in term of Top-1 beam prediction accuracy for 1600ms prediction time.

- With non-AI baseline (Option 2), similar prediction accuracy (74.1% of Top-1 beam prediction accuracy) can be achieved with 640ms prediction time.

- Under the assumption of **setting Case B+,** based on the evaluation from 1 source decent beam prediction accuracy can be achieved with 80% measurement/RS overhead comparing the non-AI baseline (Option 1, with 100% prediction accuracy) with Tper =160ms to 960ms as minimal periodicity of measurement

- evaluation results from 1 source show that AI/ML can provide 80% RS/measurement overhead reduction:

- AI/ML can achieve 68%~77% beam prediction accuracy in terms of Top-1 beam prediction accuracy with 640ms to 3840ms prediction time (4 prediction time instance) /800ms to 4800ms measurement periodicity (4 measurement time instance) without UE rotation.

##### 6.3.2.2.2 Performance when Set B is a subset of Set A

**For BM-Case2**, when *Set B patten is a subset of Set A* in each time instance, for DL Tx beam prediction with the measurements from the best Rx beam or Tx-Rx beam pair prediction, without considering generalization aspects, with the following assumptions:

- UE speed: 30km/h (unless otherwise stated)

- Prediction time: 40ms/80ms/160ms/320ms/640ms/others

- With and without UE rotation

- Fixed Set B patterns or preconfigured Set B pattens in each measurement instances (unless otherwise stated)

Note that ideal measurements are assumed:

- Beams could be measured regardless of their SNR.

- No measurement error.

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

Note: In some evaluations results, non-AI baseline (Option 2) may have better performance in terms of Top-1 beam prediction accuracy than the ratio of Set B/Set A. This is because the Top-1 beam distribution among Set A of beams are not uniform while the Set B pattern may be well designed or happen to be the beams that have high probability to be the Top-1 beam.

Note: non-AI baseline Option 2: sample and hold based on the measurements in the last time instance (unless otherwise stated)

**(A) For Tx DL beam prediction without UE rotation**, AI/ML can provide good beam prediction accuracy and gain comparing with non-AI baseline (Option 2) with same RS/measurement overhead:

- With measurements of **fixed Set B** **or variable Set B with pre-configured patterns** of beams that of **1/2** of Set A of beams in one time instance,

**- 1/2 RS overhead** in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 86.4%/83.5% prediction accuracy for prediction time 40ms/160ms,with 32 Tx beam in Set A, and Set B is different in each time instance.

- wherein, measurements from 3 time instances with measurement periodicity of 80ms are used.

- wherein, 80.5%/70% prediction accuracy can be achieved by non-AI baseline (Option 2) with assumption that the selection of 1/2 of beams selected in baseline are the most frequently used in the evaluated scenario.

- evaluation results from 1 source show that AI/ML can achieve 94.5%/93.7%/92.1% prediction accuracy for prediction time 80ms/160ms/320ms with 32 Tx beam in Set A, and Set B is the same in each time instance.

- wherein, measurements from 2 time instances with measurement periodicity of 80ms are used

- wherein, 71%/69.9%/68% prediction accuracy can be achieved by non-AI baseline with the assumption that 16 Tx beams are measured in total and preferred beam pattern is used.

- where the Rx beam of best beam pair within Set A is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 67.1%/65.01% prediction accuracy for prediction time 80ms with 32 Tx beam in Set A for 30km/h/60km/h respectively, and Set B is the same in each time instance.

- wherein, measurements from 5 time instances with measurement periodicity of 80ms are used

- wherein, 44.35%/44.29% prediction accuracy can be achieved for 30km/h/60km/h respectively by non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 75.34% prediction accuracy for prediction time 160ms with 32 Tx beams in Set A for 30km/h, and Set B is the same in each time instance.

- wherein, measurements from 4 time instances with measurement periodicity of 160ms are used

- wherein, 44.36%prediction accuracy can be achieved for 30km/h by non-AI baseline (Option 2).

- With measurements of fixed Set B or variable Set B with pre-configured patterns of beams that of 1/4 of Set A of beams in one time instance,

**- 1/4 RS overhead** in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 93.4%/92.4%/90.5% and 91.3%/90.6%/89.1% prediction accuracy for prediction time 80ms/160ms/320ms, with 32 Tx beam in Set A, and Set B is different and same in each time instance respectively

- wherein, measurements from 2 instances with measurement periodicity of 80ms are used respectively.

- Wherein, 70.5%/69.4%/67.4% and 42.5%/42.2%/41.5% prediction accuracy can be achieved by non-AI baseline (Option 2) with the assumption that 16 Tx beams are measured in total and preferred beam pattern is used.

- Where the Rx beam of best beam pair within Set A is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 56.4%/52.7% prediction accuracy for prediction time 80ms/160ms, with 64 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 2 time instances with measurement periodicity of 80ms/160ms are used respectively

- wherein, 63.25%/58.45% prediction accuracy can be achieved by non-AI baseline (Option 1) when measuring Set A during observation and then applying sample-and-hold

- evaluation results from 1 source show that AI/ML can achieve 83.15%/79.53%/79.43% prediction accuracy for prediction time 40ms/80ms/160ms, with 32 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used,

- 32.8%/32.8%/32.7% prediction accuracy can be achieved by non-AI baseline (Option 2)

- Wherein, the Rx beam of best beam pair within Set A is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 88%~90% prediction accuracy for prediction time 160ms/320ms/480ms/640ms/800ms, with 32 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 5 time instances with measurement periodicity of 160ms are used,

- 16%~22% prediction accuracy can be achieved by non-AI baseline (Option 2)

- Where the best Rx beam for each Tx beam within Set B is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 88%/86%/ 82% prediction accuracy for prediction time40ms/160ms/320ms, with 32 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 8 time instances with measurement periodicity of 40ms are used,

- 36.2%/35.8%/35.3% prediction accuracy can be achieved by non-AI baseline (Option 2) on the best Tx beam with highest L1-RSRP in the all time instances

- for random Set B pattern (Set B/Set A=1/4，the SetB is randomly changed in Set A in each time instance), compared to the above case, for Top-1 beam prediction accuracy, evaluation results show about 6% beam prediction accuracy degradation.

- wherein, the Rx beam of best beam pair within Set B is assumed to obtained the measurement of Set B

- evaluation results from 1 source show that AI/ML can achieve 73.8%/73.3% and 76.9%/73.08% prediction accuracy for prediction time 160ms/320ms, with 32 Tx beam in Set A, and Set B is the same and different in each time instance respectively

- wherein, measurements from 4 time instances with measurement periodicity of 160ms/320ms are used respectively,

- 24%/24.7% and 18.1%/17% prediction accuracy can be achieved for same and different Set B pattern respectively with non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 61.9%/56.35% prediction accuracy for prediction time 80ms with 32 Tx beam in Set A for 30km/h/60km/h respectively, and Set B is the same in each time instance.

- wherein, measurements from 5 time instances with measurement periodicity of 80ms are used

- wherein, 20.3%/22% prediction accuracy can be achieved for 30km/h/60km/h respectively by non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 61.7%~55.6% prediction accuracy for prediction time 80ms~960ms, with 32 Tx beam in Set A, and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity of equal to or 2 times of the prediction time are used respectively,

- 18.6%~8.8% prediction accuracy can be achieved for same Set B pattern with non-AI baseline (Option 2) based on the measurements of the last time instance

- Note: RS overhead reduction

- Under the assumption of setting Case A, AI/ML can achieve 57.8%~61.0% beam prediction accuracy in terms of Top-1 beam prediction accuracy for 160ms to 960ms prediction time

- up to 4/5 RS/measurement overhead reduction can be obtained with measurements from 4 time instances with measurement periodicity of 160ms to 960ms.

- Under the assumption of setting Case B, AI/ML can provide more than 90% RS/measurement overhead reduction:

- AI/ML can achieve 58% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 10% beam prediction accuracy in term of Top-1 beam prediction accuracy for 960ms prediction time

- with non-AI baseline (Option 2), 18.6% of Top-1 beam prediction accuracy can be achieved with 80ms prediction time.

- Under the assumption of setting Case B+, AI/ML can provide 87.5% RS/measurement overhead reduction:

- AI/ML can achieve 55.6%~59.5% beam prediction accuracy in terms of Top-1 beam prediction accuracy with 160ms to 960ms prediction time 320ms to 1920ms measurement periodicity (4 measurement time instance).

- evaluation results from 1 source show that AI/ML can achieve 67.25% prediction accuracy for prediction time 160ms with 32 Tx beams in Set A for 30km/h, and Set B is the same in each time instance.

- wherein, measurements from 4 time instances with measurement periodicity of 160ms are used

- wherein, 23.95% prediction accuracy can be achieved for 30km/h by non-AI baseline (Option 2).

- With measurements of fixed Set B or variable Set B with pre-configured patterns of beams that of 1/8 of Set A of beams in one time instance,

**- 1/8 RS overhead** in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 67.4%/67.8%/ 70%/66.9%/67.5%/64.9%/62.9% prediction accuracy for prediction time 160ms/320ms/480ms/ 640ms/800ms/960ms, with 32 Tx beam in Set A, and Set B is the same in each time instance.

- wherein, measurements from 8 time instances with measurement periodicity of 160ms are used

- 9%/8.9%/8.8%/8.7%/8.5%/8.4% prediction accuracy can be achieved by non-AI scheme (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 94%/93.5%/92.6%/90.7% prediction accuracy for prediction time 40ms/80ms/160ms/320ms, with 32 Tx beam in Set A, and Set B is different in each time instance respectively

- wherein, measurements from 4 time instances with measurement periodicity of 40ms is used.

- wherein, 70.7%/70.2%/69.1%/67.2% prediction accuracy can be achieved by non-AI baseline (Option 2) with the assumption that 16 Tx beams are measured in total and preferred beam pattern is used.

- where the Rx beam of best beam pair within Set A is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 76.1%/75.2%/70.7% prediction accuracy for prediction time 40ms/80ms/160ms, with 32 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used,

- 18.0%/17.9%/17.8% prediction accuracy can be achieved by non-AI baseline (Option 2)

- wherein the Rx beam of best beam pair within Set A is assumed to obtained the measurement of Set B.

- evaluation results from 1 source show that AI/ML can achieve 81.7%/81.1%/80.6% prediction accuracy for prediction time 40ms/160ms/320ms, with 32 Tx beam in Set A and Set B is the same in each time instance

- wherein, measurements from 8 time instances with measurement periodicity of 40ms are used,

- 30.7%/30.4%/30% prediction accuracy can be achieved by non-AI baseline (Option 2) based on the best Tx beam with highest L1-RSRP in all the time instances

- for random Set B pattern (SetB/SetA=1/8，the SetB is randomly changed in Set A in each time instance), compared to the above case, for Top-1 beam prediction accuracy, evaluation results show about 5% beam prediction accuracy degradation.

- wherein, the Rx beam of best beam pair within Set B is assumed to obtained the measurement of Set B

- evaluation results from 1 source show that AI/ML can achieve 56.91% prediction accuracy for prediction time 160ms with 32 Tx beams in Set A for 30km/h, and Set B is the same in each time instance.

- wherein, measurements from 4 time instances with measurement periodicity of 160ms are used

- wherein, 18.75% prediction accuracy can be achieved for 30km/h by non-AI baseline (Option 2).

**(B) For Tx DL beam prediction with UE rotation**, based on evaluation from 2 sources, AI/ML can provide good beam prediction accuracy and gain comparing with non-AI baseline (Option 2) with same RS/measurement:

- With measurements of **fixed Set B** of beams that of **1/3** of Set A of beams in one time instance. (Note that more RS overhead can be achieved considering additional temporal domain RS overhead reduction)

**- 1/3 RS overhead** in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Evaluation results from 1 source show that AI/ML can achieve

- 77.5% Top-1 beam prediction accuracy for 160ms prediction time and 200ms measurement periodicity wherein, 33.4% prediction accuracy can be achieved by non-AI baseline (Option 2), and 43.3% beam prediction accuracy can be achieved by a combination of spatial interpolation (radial basis function interpolation) followed by sample-and-hold.

- Under the assumption of Case B+, 93.3% RS overhead reduction can be achieved compared to non-AI baseline (Option 1) assuming all Set A of beams needs to be measured every 40ms at each time instances for measurement and prediction.

- Wherein, UE rotation is modelled every 40ms with constant 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes.

- With measurements of **variable Set B** (with preconfigured Set B pattern in each time instances) of beams that of **1/3** of Set A of beams in one time instance,

**- 1/3 RS overhead** in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Evaluation results from 1 source show that AI/ML can achieve

- 78%/76%/73.8%/68.6% Top-1 beam prediction accuracy for 160ms/320ms/480ms/960ms prediction time and 200ms/360ms/520ms/1000ms measurement periodicity

- wherein, 71.5%/63%/56.5%/45.3% prediction accuracy can be achieved by non-AI baseline (Option 2), in which for each prediction instance, the latest measurement for each beam in Set A is used as the predicted value for that beam.

- wherein, Set B patterns in Set A/Set B consecutive time slots partition Set A.

- Under the assumption of Case B+, **93.3%/96.3%/97.4%/98.7% RS overhead reduction** can be achieved compared to non-AI baseline (Option 1) assuming all Set A of beams needs to be measured every 40ms at each time instances for measurement and prediction for 160ms/320ms/480ms/960ms prediction time.

- Wherein, UE rotation is modelled every 40ms with constant 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes.

- With measurements of **fixed Set B** of beams that of **1/4** of Set A of beams in one time instance,

- 1/4 RS overhead in spatial domain can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 DL Tx beam prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 71.8%/57.3% prediction accuracy for prediction time 160ms/320ms, with 32 Tx beam in Set A, and Set B is the same in each time instance respectively

- wherein, measurements from 4 time instances with measurement periodicity of 160ms/320ms are used respectively,

- 24.3%/14.2% prediction accuracy can be achieved for same and different Set B pattern respectively with non-AI baseline (Option 2)

- Wherein, UE rotation is modelled every 20ms with a rotation speed uniformly distributed within {0, 60} RPM, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

**(C) For beam pair prediction without UE rotation**, based on evaluation of most sources, AI/ML can provide good beam prediction accuracy and gain comparing with non-AI baseline (Option 2) with same RS/measurement overhead.

- With measurements of fixed Set B or variable Set B with preconfigured pattern in each time instance of beams that of **1/4** of Set A of beams in one time instance,

**- 1/4 RS overhead** can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 beam pair prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 76.3%/74.7%/72% prediction accuracy for prediction time 40ms/80ms/160ms, with 32 Tx beams and 8 Rx beams in Set A, and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used

- 32.7%/32.6%/32.5% prediction accuracy can be achieved by non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 88%~90% prediction accuracy for prediction time 160ms/320ms/480ms/640ms/800ms, with 32 Tx beams and 8 Rx beams in Set A, and Set B is the same in each time instance

- wherein, measurements from 5 time instances with measurement periodicity of 160ms are used

- 19%~23% prediction accuracy can be achieved by non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 80.97%/80.17%/75.86% prediction accuracy for prediction time 40ms/80ms/160ms, with 32 Tx beam and 4 Rx beam in Set A, and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used,

- 38.6%/38.0%/37.2% prediction accuracy can be achieved by non-AI baseline (Option 2)

- evaluation results from 1 source show that AI/ML can achieve 63.2%/~57.7% prediction accuracy for prediction time 80ms~960ms, with 32 Tx beam and 8 Rx beam in Set A, and Set B is the same in each time instance

- wherein, measurements from 4 time instances with measurement periodicity same as or 2 times of the prediction time are used

- 22.3%~10.7% prediction accuracy can be achieved by non-AI baseline (Option 2)

- RS overhead reduction

- Under the assumption of setting Case A, AI/ML can achieve 58.1%~62.0% beam prediction accuracy in terms of Top-1 beam prediction accuracy for 160ms to 960ms prediction time, up to 4/5 RS/measurement overhead reduction can be obtained with measurements from 4 time instances with measurement periodicity of 160ms to 960ms.

- Under the assumption of setting Case B, AI/ML can provide more than 90% RS/measurement overhead reduction:

- AI/ML can achieve 58.1% beam prediction accuracy, while non-AI baseline (Option 2) can only achieve 12.7% beam prediction accuracy in term of Top-1 beam prediction accuracy for 960ms prediction time

- With non-AI baseline (Option 2), 22.3% of Top-1 beam prediction accuracy can be achieved with 80ms prediction time.

- Under the assumption of setting Case B+, AI/ML can provide 87.5% RS/measurement overhead reduction:

- AI/ML can achieve 57.1%~60.7% beam prediction accuracy in terms of Top-1 beam prediction accuracy with 160ms to 960ms prediction time /320ms to 1920ms measurement periodicity (4 measurement time instance).

- evaluation results from 1 source show that AI/ML can achieve 48.2%/51.6% prediction accuracy for prediction time 160ms, with 32 Tx beam and 8 Rx beam in Set A, and Set B is the same and different in each time instance respectively

- wherein, measurements from 4 time instances with measurement periodicity of 160ms are used,

- 16.2%/22.9% prediction accuracy can be achieved by non-AI baseline (Option 2) based on the measurements of the last time instance

- With measurements of fixed Set B of beams that of **1/8** of Set A of beams in one time instance,

**- 1/8 RS overhead** can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 beam pair prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 76.7%/74.1%/73.6% prediction accuracy for prediction time 40ms/160ms/320ms, with 256 (32Tx\*8Rx) beam pairs in Set A and Set B (4Tx\*8Rx) is the same in each time instance

- wherein, measurements from 8 time instances with measurement periodicity of 40ms are used,

- 30.1%/29.7%/29.1% prediction accuracy can be achieved by non-AI baseline (Option 2) based on the measurements in all time instances

- evaluation results from 1 source show that AI/ML can achieve 77.0%/76.2%/72.0% and 74.2%/73.0%/69.8% prediction accuracy for prediction time 40ms/80ms/160ms, with 32 Tx beams and 4 Rx beams in Set A, and Set B is the same in each time instance with all measurements from all Rx beams and half of Rx beams respectively

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used,

- 9.88%/9.60%/8.95% and 14.57%/14.45%/14.27% prediction accuracy can be achieved by non-AI baseline (Option 2) for the case with all Rx beams and half of Rx beams respectively

- With measurements of fixed Set B or variable Set B with pre-configured pattern in each time instance of beams that of **1/16** of Set A of beams in one time instance,

**- 1/16 RS overhead** can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 beam pair prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 50.58%/48.71%/44.33% and 63.94%/63.31%/60.49% prediction accuracy for 40ms/80ms/160ms prediction time with 32 Tx beam in Set A, and Set B is the same in each time instance with {8 Tx and 2 Rx} and {4 Tx and all Rx} respectively.

- wherein, measurements from 4 time instances with measurement periodicity of 40ms are used

- 8.96%/8.91%/8.89% and 4.7%/4.56%/4.3% prediction accuracy can be achieved by non-AI scheme (Option 2) for the case with from all Rx beams and half of Rx beams respectively

- evaluation results from 1 source show that AI/ML can achieve 89.1% / 86.4%/ 82.9% prediction accuracy for prediction time 40ms/160ms/320ms, with 256 (32Tx\*8Rx) beam pairs in Set A and Set B (2Tx\*8Rx) is different in each time instance

- wherein, measurements from 8 time instances with measurement periodicity of 40ms are used,

- 69.4%/67.8%/66% prediction accuracy can be achieved by non-AI baseline (Option 2) based on the measurements in all time instances

**(D) For beam pair prediction with UE rotation**, evaluations from 2 sources show AI/ML can provide 44% or 15% beam prediction accuracy gain comparing with non-AI baseline (Option 2) with same RS/measurement overhead, with 78% or 30%~35% Top-1 beam prediction accuracy respectively.

- With measurements of fixed Set B or variable Set B with pre-configured pattern in each time instance of beams that of **1/4** of Set A of beams in one time instance,

**- 1/4 RS overhead** can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 beam pair prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 35.02%/29.2% prediction accuracy for prediction time 40ms/160ms, with 32 Tx beam and 8 Rx beam in Set A, and Set B is the same and different in each time instance respectively

- wherein, measurements from 4 time instances with measurement periodicity of 40ms/160ms are used,

- 19.7%/15.6% prediction accuracy can be achieved by non-AI baseline (Option 2) based on the measurements of the last time instance

- UE rotation is modelled every 20ms with a rotation speed uniformly distributed within {0, 60} RPM, and the rotation direction is {1/4 of data with randomly to left or right in horizontal, 1/4 of data always to left, 1/4 of data always to right, 1/4 of data to left and right in turn} with random initial directly.

- With measurements of variable Set B with pre-configured patterns in each time instance of beams that of **1/16** of Set A of beams in one time instance,

**- 1/16 RS overhead** can be achieved comparing with non-AI baseline (Option 1) assuming all Set A of beams needs to be measured at each time instances for measurement and prediction. More RS overhead can be achieved considering additional temporal domain RS overhead reduction.

- Top-1 beam pair prediction accuracy:

- evaluation results from 1 source show that AI/ML can achieve 78.1% prediction accuracy for prediction time 40ms with 32 Tx beams and 8 Rx beams in Set A, Set B is different in each time instance and 10 RPM rotation speed to fixed a direction

- wherein, measurements from 3 time instances with measurement periodicity of 40ms or 80ms are used

- 42.4%/42.5% prediction accuracy can be achieved by non-AI scheme (Option 2).

#### 6.3.2.3 Performance under different assumptions/scenarios for BM-Case1 and/or BM-Case2

**Performance with quantization:**

At least for BM-Case1 for inference of DL Tx beam with L1-RSRPs of all beams in Set B, existing quantization granularity of L1-RSRP (i.e., 1 dB for the best beam, 2 dB for the difference to the best beam) causes a minor loss in beam prediction accuracy compared to unquantized L1-RSRPs of beams in Set B.

- Evaluation results from 13 sources show less than 5% beam prediction accuracy degradation in terms of Top-1 beam prediction accuracy.

- Note: 1 source uses the data without quantization for training and data with quantization for inference. Other sources use the same quantization scheme for data for training and inference.

At least for BM-Case1 for inference of DL Tx beam with L1-RSRPs of all beams in Set B,

- Evaluation results from 4 sources show that, with 1dB quantization step for the absolute L1-RSRP of the best beam and 4dB quantization step differential L1-RSRP report with the existing quantization range, less than 5% beam prediction accuracy degradation in terms of Top-1 beam prediction accuracy compared to unquantized L1-RSRPs of beams in Set B.

- Same quantization scheme is used for the input data for training and inference.

- Note: 1 source used quantized L1-RSRPs with the same quantization scheme as labels in training.

- Note: 1 source used unquantized L1-RSRPs as labels in training.

- Note: 1 source used unquantized L1-RSRPs to determine Top-1 beam id as labels in training.

**Performance with measurement error**

**For BM-Case1 DL Tx beam prediction** (unless otherwise stated), when *Set B is a subset* (1/4 unless otherwise stated) *of Set A*, **without differentiating BB errors and RF errors** modelled as truncated Gaussian distribution (unless otherwise stated),

- Considering ±2 dB relative measurement error,

- evaluation results from 3 sources show that the beam prediction accuracy degrades 6%~10%in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And 1 source shows that 95%ile of L1-RSRP diff can be about 1.4~2dB, 1 source shows that average L1-RSRP diff can be lower than 1dB.

- evaluation results from 1 source show that

- for DL Tx beam prediction, the beam prediction accuracy degrades 28.8% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error, [and average L1-RSRP diff can be about 7.3dB.

- for Tx-Rx beam pair prediction when Set B is 1/8 of Set A, the beam prediction accuracy degrades 2.4% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error, and average L1-RSRP diff can be about 5.8dB

- wherein the measurement error is modelled as uniformed distribution.

- evaluation results from 1 source show that considering different relative measurement error range in model training (±2 dB, ±0 dB), similar (less than 1% difference) Top-1 beam prediction accuracy can be achieved

- Considering ±3 or ±4 dB relative measurement error,

- evaluation results from 4 sources show that the beam prediction accuracy degrades 14% (with 3dB error) ~20% (with 4dB error) in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And 1 source shows that the 95%ile of L1-RSRP diff can be about 2~3.2dB. 1 source shows that average L1-RSRP diff can be lower than 1dB.

- evaluation results from 1 source show that considering different relative measurement error range in model training (0dB, ±2 dB, ±4 dB), similar (less than 1% difference) Top-1 beam prediction accuracy can be achieved, and average L1-RSRP diff can be lower than 1dB when ±2 dB or ±4 dB relative measurement error is considered in model training

- Considering up to ±5 dB relative measurement error when Set B is 1/8 of Set A,

- evaluation results from 1 source show that the beam prediction accuracy degrades 13.6% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error for DL Tx beam prediction.

- Considering ±6 dB relative measurement error,

- evaluation results from 3 sources show that the beam prediction accuracy degrades 22%~30% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And the 95%ile of L1-RSRP diff can be about 3.1~7.5dB.

- evaluation results from 1 source show that he L1-RSRP difference in 90%ile degrades 7dB for the AI/ML model, compared to baseline 1 and 2 that degrades 3 dB respectively 1 dB at the same percentile.

- evaluation results from 1 source show that for both DL Tx beam prediction and beam pair prediction, the beam prediction accuracy degrades 42~48% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And the average L1-RSRP diff can be about 1.6dB.

- However, comparing with the global search of all beams in Set A with the same measurement error level, for DL Tx beam prediction the beam prediction accuracy degrades less than 1% in terms of Top-1 beam prediction accuracy, and for Tx-Rx beam pair prediction the beam prediction accuracy degrades about 7% in terms of Top-1 beam prediction accuracy.

- Note: in this evaluation, measurement errors are considered in training and inference phase only for AI inputs with idea labels in training phase.

- evaluation results from 1 source show that

- for DL Tx beam prediction, the beam prediction accuracy degrades 32.4% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error, [and average L1-RSRP diff can be about 8.34dB.

- for Tx-Rx beam pair prediction, the beam prediction accuracy degrades 5.2% in terms of Top-1 beam prediction accuracy comparing to the one without measurement error, [and average L1-RSRP diff can be about 6.4dB.

- evaluation results from 1 source show that considering different relative measurement error range in model training (0dB, ±2 dB, ±6 dB), similar less or than 2% Top-1 beam prediction accuracy can be achieved, and average L1-RSRP diff can be lower than 1dB when ±6 dB relative measurement error is considered in model training

**For BM-Case1 DL Tx beam prediction or Tx-Rx beam pair prediction**, when *Set B is a subset* (1/4 unless otherwise stated) *of Set A***, with separately modelled BB error and/or RF errors** modelled as truncated Gaussian distribution (unless otherwise stated),

- Considering ±3 relative measurement error for BB and RF respectively,

- evaluation results from 1 source show that for DL Tx beam prediction and beam pair prediction with Set B is ¼ of Set A, the beam prediction accuracy degrades 42% and 38% respectively in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And the average of L1-RSRP diff is about 1.1dB and 2.16dB respectively.

- However, comparing with the global search of all beams in Set A with the same measurement error level, for DL Tx beam prediction the beam prediction accuracy degrades about 2 % in terms of Top-1 beam prediction accuracy, and for Tx-Rx beam pair prediction the beam prediction accuracy degrades about 8% in terms of Top-1 beam prediction accuracy.

- Note: in this evaluation, measurement errors are considered in training and inference phase only for AI inputs with idea labels in training phase.

- evaluation results from 1 source show that for both DL Tx beam prediction with Set B is 1/4 of Set A and beam pair prediction with Set B is 1/16 Set A, the beam prediction accuracy degrades 4.3% and 6.3% respectively in terms of Top-1 beam prediction accuracy comparing to the one without measurement error. And the average of L1-RSRP diff becomes 0.7dB and 2.18dB larger respectively.

- Note: in this evaluation, for DL Tx beam prediction, the measurements of Set B from each Rx beam of all Rx beams were used as AI inputs to obtain Top-K beams, followed by Top-K beam sweeping with that given Rx beam. This procedure repeats over all Rx beams, to obtain the best Tx beam at all Rx beams.

- Considering 3.3 dB for standard deviation in relative measurement error without truncation for RF only, evaluations results from 1 source show with AI/ML:

- with a common measurement error for all Tx beams at a given Rx beam:

- Top-1 beam prediction accuracy with 1 dB margin performance has slight performance degradation (less than 0.2%) than that without measurement error.

- with independent measurement errors for all Tx beams,

- Top-1 beam prediction accuracy with 1 dB margin has 10% and 20% performance degradation than that without measurement error for Set B/Set A = 1/2 and 1/4 respectively.

- wherein, measurement errors are only considered in inference inputs

Note that:

- In the above results, measurement errors are considered in both training (input data and label) and inference phase (except the ground-truth) unless otherwise stated.

- Beams could be measured regardless of their SNR.

- Measured in a single-time instance (within a channel-coherence time interval).

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

**Performance with different Rx beam assumption for DL Tx beam prediction**

At least for BM-Case1 when Set B is a subset of Set A, and for **DL Tx beam prediction**, with the measurements of the "best" Rx beam with exhaustive beam sweeping for each model input sample, AI/ML provides the better performance than with measurements of random Rx beam(s).

- Evaluation results from 12 sources show 20%~50% degradation with random Rx beam(s) comparing with the "best" Rx beam in terms of Top-1 prediction accuracy.

- Evaluation results from 1 source shows 12% degradation with measurement of random Rx compared with measurement of best Rx in term of Top-1 beam prediction accuracy.

Comparing performance with non-AI baseline option 2 (based on the measurement from Set B of beams), with measurements of random Rx beam(s) as AI/ML inputs:

- Evaluation results from 7 sources show that AI/ML can still provide 7%~44% beam prediction accuracy gain in terms of Top-1 beam prediction accuracy.

Note: In both training and inference, measurements of random Rx beams are used as AI/ML inputs.

**For BM-Case 1 DL Tx beam prediction without UE rotation**, for Top-1 beam prediction accuracy, compared to the best Rx beams obtained from one shot measurements, with quasi-optimal Rx beam performance degradation is observed:

- evaluation results from 1 source show 2% beam prediction accuracy degradation when Set B = 1/2 Set A and 7% beam prediction accuracy improvement when Set B = 1/4 or 1/8 Set A, when using the best Rx beams obtained from previous exhaustive sweeping (20ms ago) of all beams in Set A, comparing with using the best Rx beam for each Tx beams in Set B obtained from current exhaustive sweeping, without considering UE rotation for 3km/h UE speed. Such beam prediction accuracy improvement may not exist when considering UE rotation and higher UE speed.

- evaluation results from 1 source show 2.5% beam prediction accuracy degradation using the best Rx of each Tx beams obtained from previous exhaustive sweeping (20ms ago) than using the best Rx of each Tx beams obtained from current exhaustive sweeping, without considering UE rotation for 3km/h UE speed.

- evaluation results from 1 source shows 6.6%/6.9%/32.1%/45% degradation using a stochastic model in which the UE Rx beam is randomly selected with average probability that the best Rx beam is selected equal to 87.1%/75.1%/34.3%/10.9% compared to using the best Rx of each Tx beams obtained from current exhaustive sweeping, without considering UE rotation

- evaluation results from 1 source show 13% beam prediction accuracy degradation, with the assumption of the best Rx beam for each Tx beam obtained from previous exhaustive sweeping over all beams in Set A in a SSB-like structure (in the past 160ms for each Rx beam with every 20ms a burst of Set A of beams) without considering UE rotation for 3km/h UE speed.

- evaluation results from 1 source show 3%~11% beam prediction accuracy degradation, with the assumption of the best Rx beam obtained from one specific Tx beam which is 1st Tx beam in Set B.

- evaluation results from 1 source show 12% beam prediction accuracy degradation with the assumption of the best Rx beams obtained from one specific Rx beam which is the best between the same Rx beam for different panels.

- In addition, evaluation results from 3 sources show 1%~4% and 6%~12% beam prediction accuracy degradation, with the assumption of the best Rx beam is used for 90% and 80% of the model input samples and random Rx beam for the remaining samples respectively.

- Even though, AI/ML can still provide better performance than non-AI baseline option 2 (exhaustive beam sweeping in Set B of beams), e.g., 50%~60% beam prediction accuracy difference in terms of Top-1 beam prediction accuracy based on the evaluation results from 2 sources, where non-AI baseline option 1 (exhaustive beam sweeping in Set A of beams) provides 100% prediction accuracy.

**For BM-Case 2 DL Tx beam prediction with UE rotation**, for Top-1 beam prediction accuracy, with quasi-optimal Rx beam selection:

- evaluation results from 1 source show 5~11% beam prediction accuracy improvement given the assumption of the best Rx beams obtained from previous round-robin sweep of beam pair links from beams in Set A, compared to sample-and-hold baselines.

- In the evaluation, UE rotation is modelled every 40ms with constant 10 RPM rotation speed in all three rotational axes, with rotational direction chosen uniformly at random among the three axes.

**Performance with different label options**

Different label options may lead to different data collection overhead for training. At least for BMCase-1, for (Option 1a) Top-1 beam(pair) in Set A as the label and (Option 2a) all L1-RSRPs per beam of all the beams(pairs) in Set A as the label, with the comparable model complexity and computation complexity, the results across companies and the observed performance delta are summarized as below:

- For Top 1 beam (pair) prediction accuracy,

- evaluation results from 7 sources show that an AI/ML model with Top-1 beam(pair) in Set A as the label (Option 1a) can provide better performance (e,g, 2~7% or 12%~18% higher for Top 1 beam prediction accuracy) than an AI/ML model with all L1-RSRPs per beam of all the beams(pairs) in Set A as the label (Option 2a)

- evaluation results from 1 source show that similar or slightly worse (e,g, 2% higher for Top 1 beam prediction accuracy)) can be achieved with Option 1a than Option 2a

- For Top-K beam (pair) prediction accuracy or Top-1 beam prediction accuracy with 1dB margin,

- evaluation results from 2 sources show that Option 1a can provide similar performance than Option 2a

- evaluation results from 1 source show that Option 2a can provide 5%~12% better performance than Option 1a for Top-2/-4 beam pair prediction accuracy.

- evaluation results from 1 source show that show that Option 1a can provide 2%~5% better performance than Option 2a for Top-2/-6 beam pair prediction accuracy.

- evaluation results from 1 source show that show that Option 1a can provide 2%~7% /1%~5% better performance than Option 2a for Top-2/-4 beam prediction accuracy for DL Tx beam prediction.

- evaluation results from 1 source show that show that Option 1a can provide <1% or 9%~17% better performance than Option 2a for Top-2/-3 beam prediction accuracy for DL Tx beam prediction for Set B=1/2 Set A or Set B =1/4 or 1/8 Set A.

- Detailed assumptions and results are listed as below:

- evaluation results from one source show that for both DL Tx beam prediction and beam pair prediction with Set B is ¼ of Set A, with Top-1 beam in Set A as the label, AI/ML can provide 2%~3% higher beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity. The Top-K beam prediction accuracy is comparable for DL Tx beam prediction; however, the Top-K beam prediction accuracy is slightly better (<1%) with all L1-RSRPs as the label. The average L1-RSRP difference is similar (about 1.5dB) in the two cases.

- evaluation results from one source show that for Tx beam prediction with Set B is 1/2 Set A and Set B is 1/4 Set A, with Top-1 beam in Set A as the label, AI/ML can provide 2%-5% higher beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity. The Top- 1 beam with 1dB error and Top-K beam prediction accuracy is comparable for DL Tx beam prediction.

- evaluation results from one source show that for beam pair prediction with Set B is 1/8 or 1/16of Set A, with Top-1 beam in Set A as the label, AI/ML can provide 4%-6% higher beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label even with larger model complexity.

- evaluation results from one source show that for beam pair prediction with Set B is ¼ Set A, with Top-1 beam in Set A as the label, AI/ML can provide 12% higher beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs of all the beams as the label with comparable model complexity. However, labelling with all L1-RSRPs can provide 5% and 12 % better for Top-3 or Top-4 beam prediction accuracy comparing with labelling with Top-1 beam ID.

- evaluation results from one source show that for beam pair prediction with Set B is ¼ Set A, with Top-1 beam in Set A as the label, AI/ML can provide 15% higher beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity. The average L1-RSRP difference is similar (about 0.4dB) in the two cases.

- evaluation results from one source show that for DL Tx beam prediction with Set B is ¼ of Set A, with Top-1 beam in Set A as the label, AI/ML can provide similar beam prediction accuracy in terms of Top-1 beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label. Using Top-1 beam as the label can provide 2%/5% better performance for Top-2/-6 beam prediction. The average L1-RSRP difference is similar (about 1dB) in the two cases.

- evaluation results from one source show that for beam pair prediction with Set B is 1/16 of Set A, with Top-1 beam in Set A as the label, 2% beam prediction accuracy degradation in terms of Top-1 beam prediction accuracy is achieved comparing to the one with all L1-RSRPs per beam of all the beams as the label.

- evaluation results from one source show that for Tx beam prediction with Set B is 1/4 of Set A or 1/8 of Set A or 1/16 of Set A, with Top-1 beam in Set A as the label, AI/ML can provide comparable or up to 7% higher beam prediction accuracy in terms of Top-K (K=1, 2, 4) beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity. However, the performance of average L1-RSRP difference of Top-1 predicted beam and beam prediction accuracy with 1dB margin for Top-1 beam is comparable or better with all L1-RSRPs per beam of all the beams as the label.

- Evaluation results from one source show that for Tx beam prediction with Set B is 1/2 Set A, with Top-1 beam in Set A as the label, AI/ML can provide <1% higher beam prediction accuracy in terms of Top-K (K=1,2,3) beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity. With Set B is 1/4 Set A and 1/8 Set A and Top-1 beam in Set A as the label, AI/ML can provide 10-18% higher beam prediction accuracy in terms of Top-K (K=1,2,3) beam prediction accuracy comparing to the one with all L1-RSRPs per beam of all the beams as the label with comparable model complexity.

In addition, 1 source show good performance with Top-K beam(pair)s in Set A and the corresponding L1-RSRPs as the label (Option 2b) can be achieved with two separate AI models. In the evaluation, one classification model (with Top-1/K beam(s) in Set A as the label(s)) is used to predict the Top-1/K beam and another regression model (with L1-RSRP(s) of Top-1/K beam(s) in Set A as the label(s)) is used to predict L1-RSRP(s).

Note: The performance for beam predication accuracy with AI/ML may also depend on some other aspects, e.g., AI/ML model architecture choice, model training parameters (e.g., hyperparameter tuning), loss function corresponding to optimizing certain KPI(s). Assumptions on loss function are not indicated in the evaluations above.

Note: ideal measurements are assumed

- Beams could be measured regardless of their SNR.

- No measurement error.

- Measured in a single-time instance (within a channel-coherence time interval).

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

**Performance with different Set B pattern assumptions**

For BMCase-1 and for a fixed Set B pattern option, Set B pattern will affect the beam prediction accuracy with AI/ML for both DL Tx beam prediction and beam pair prediction.

At least for BM-Case1 (unless otherwise stated) DL Tx beam with the measurements from the best Rx beam, and/or beam pair prediction, when Set B is a subset of Set A without considering other generalization aspects and without UE rotation.

**- (Opt 2B)** For the case that Set B of beam(pair)s is changed among pre-configured patterns, compared to the case that Set B is fixed across training and inference (Opt 1), for Top-1 beam prediction accuracy

- evaluation results from 14 sources show no more than 10% or about 10% beam prediction accuracy degradation, wherein 2 sources used up to 24 pre-configured patterns and the rest of sources use 3 ~ 5 patterns;

- AI/ML still can provide better performance (e.g., >30%) of Top-1 beam prediction unless otherwise stated) than non-AI baseline option 2 (exhaustive beam sweeping in Set B of beams).

- Note: the above performance can also be treated as training with mixed patterns of Set B of beam, and testing with mixed patterns Set B of beams.

**- (Opt 2C)** For the case that Set B of beam(pair)s is randomly changed in Set A of beams, compared to the case that Set B is fixed across training and inference (Opt 1), for Top-1 beam prediction accuracy

- evaluation results from 2 sources show 10%~20% beam prediction accuracy degradation.

- evaluation results from 7 sources show 20%~50% beam prediction accuracy degradation.

- AI/ML still can provide better performance (e.g., >25% of Top-1 beam prediction unless otherwise stated) than non-AI baseline option 2 (exhaustive beam sweeping in Set B of beams):

**- (Opt 2D)** For the case that Set B of beams (pairs) is a subset of measured beams (pairs) Set C (where Set C is fixed across training and inference), compared to the case with all measurements of measured beam Set C as AI inputs

**- with Top K=1/2** of the measurements of Set C,

- For Top-1 beam prediction accuracy

- evaluation results from 5 sources show less than 4% the beam prediction accuracy degradation

- evaluation results from 3 sources show about 7% the beam prediction accuracy degradation

- evaluation results from 1 source show <1% and 7% beam prediction accuracy degradation with measuring 1/2 and 1/4 of Set A of beams respectively.

- evaluation results from 1 source show about 12% the beam prediction accuracy

- Note: all the above results are for DL Tx beam prediction

- For NW-side model, 1/2 UCI reporting overhead for inference inputs can be saved without considering quantization impact.

- In the above evaluation, 5 sources use L1-RSRPs of Top-4 measurements of 8 beams in Set C for 32 Tx beams in Set A.

- In the above evaluation, 3 sources use L1-RSRPs of Top-8 measurements of 16 beams in Set C for 64 Tx beams in Set A

- In the above evaluation, 1 source uses L1-RSRPs of Top-4/-8 measurements of 8/16 beams in Set C for 32 Tx beams in Set A.

**- with** **Top K=1/4** of the measurements of Set C,

- For Top-1 beam prediction accuracy

- evaluation results from 2 sources show 8~10% beam prediction accuracy degradation.

- evaluation results from 1 source show 15% beam prediction accuracy degradation.

- evaluation results from 1 source show 2% beam prediction accuracy degradation with measuring 1/2 of Set A of beams respectively.

- Note: all the above results are for DL Tx beam prediction

- For NW-side model, 3/4 UCI reporting overhead for inference inputs can be saved without considering quantization impact.

- In the above evaluation, 1 source uses L1-RSRPs of Top-4 measurements of 16 beams in Set C for 32 Tx beams in Set A.

- In the above evaluation, 2 sources use L1-RSRPs of Top-4 measurements of 16 beams in Set C for 64 Tx beams in Set A.

**- with** **Top K=1/8** of the measurements of Set C,

- evaluation results from 1 source show 7.5% beam prediction accuracy degradation in terms of Top-1 beam prediction accuracy for beam pair prediction.

- For NW-side model, 7/8 UCI reporting overhead for inference input can be saved.

- In the evaluation, 1 resource uses L1-RSRPs of Top-16 measurements of 128 beams in Set C for 64 Tx beams and 8 Rx beams in Set A.

**- with Top K=1/6** of the measurements of Set C, for BM-Case 2, evaluation results [from 1 source] show 3.5% improvement in beam prediction accuracy compared to non-AI/ML baseline (Option 2, sample-and-hold) whose beam prediction accuracy is 78.2%.

**- with the reported measurements** **within a given gap** of [5dB/ 10dB/ 14dB~20dB] to the best beam in Set C, evaluation results from 6 sources show 15%~28% / 4%~16.4%/ 2%~6% respectively beam prediction accuracy degradation.

- 1 source Samsung simulated for BM-Case 2, and filled in the unreported measurements in Set C as (L1-RSRP of the best Rx beam in Set C–14dB) as the inputs for AI/ML.

**- with Top-M measurements** in Set C or with the **reported measurements within a given gap** to the best beam in Set C (when Set C is larger than Set B), comparing with the case that using a smaller number of beams in Set B as the fixed pattern, the results show that comparable or better beam prediction accuracy can be achieved with the same reporting overhead or numbers of measurements as of AI inputs but larger measurement overhead.

- evaluation results from 1 source show similar Top-1 beam prediction accuracy for the case using the measurements of Top 8 beams of 16 beams in Set C and 64 beams in Set A comparing with using 8 fixed beams in Set B.

- evaluation results from 1 source show 16.5% and 43% gain in terms of Top-1 beam prediction accuracy for the case of using the measurements of Top 4 beams of 8 or 16 beams in Set C and 32 beam in Set A respectively comparing with using 4 fixed beams in Set B.

- evaluation results from 1 source show about 8% gain in terms of Top-1 beam prediction accuracy for the case using the measurements of Top 4 beams of 8 beams in Set C and 32 beams in Set A comparing with using 4 fixed beams in Set B.

- evaluation results from 1 source show about 12.5% gain in terms of Top-1 beam prediction accuracy for the case using the measurements of Top 4 beams of 8 beams in Set C and 32 beams in Set A comparing with using 4 fixed beams in Set B.

- evaluation results from 1 source show about 18% gain in terms of Top-1 beam prediction accuracy for the case using the measurements of Top 8 beams of 16 beams in Set C and 64 beams in Set A comparing with using 4 beams in Set B.

- evaluation results from 1 source show similar Top-1 beam prediction accuracy for the case using the measurements of Top 4 beams of 8 beams in Set C and 32 beams in Set A comparing with using 4 fixed beams in Set B

- evaluation results from 1 source show 17% gain in terms of Top-1 beam prediction accuracy for the case of using the measurements of Top 8 beams of 16 beams in Set C and 64 beams in Set A comparing with using 8 fixed beams in Set B. .

- evaluation results from 1 source show 12% gain in terms of Top-1 beam prediction accuracy for the case of using the measurements of Top 4 beams of 8 in Set C and 32 beam in Set A comparing with using 4 fixed beams in Set B respectively.

- The beam prediction accuracy increases with the number of measurements of Set B.

- AI/ML still can provide better performance (e.g., >30% of Top-1 beam prediction unless otherwise stated) than non-AI baseline option 2 (exhaustive beam sweeping in Set B of beams).

- Note that ideal measurements are assumed

- Beams could be measured regardless of their SNR.

- No measurement error.

- Measured in a single-time instance (within a channel-coherence time interval).

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

- This observation is based on Set B patterns that were chosen by each company.

- Implicit or explicit information of Tx beam ID and/or Rx beam ID are used as AI/ML model inputs

#### 6.3.2.4 Generalization Performance for BM-Case1 and BM-Case2

The following *generalization aspects* were evaluated for at least BMCase-1 when Set B is a subset of Set A (and BMCase-2 if stated),

- Scenarios

- Various deployment scenarios,

- e.g., UMa, UMi

- e.g., 200m ISD or 500m ISD

- Various outdoor/indoor UE distributions, e.g., 100%/0%, 20%/80%, and others

- Various UE mobility (for BMCase-2 only),

- e.g., 30km/h, 60km/h and others

- Configurations (parameters and settings)

- Various UE parameters,

- e.g., UE codebook

- e.g., UE antenna array dimensions

- e.g., different number beams in a seen UE codebook when inference using a subset of Rx beams of training

- Various gNB settings,

- e.g., DL Tx beam codebook

- e.g., gNB antenna array dimensions

- Various Set A of beam(pairs)

- Various Set B of beam (pairs)

Note: the following are assumed in the simulation unless otherwise stated

- For DL Tx beam prediction, the measurements from best Rx beam are used.

- Fixed Set B pattern.

- Without UE Rotation.

- Beams could be measured regardless of their SNR.

- No measurement error.

- Measured in a single-time instance (within a channel-coherence time interval).

- No quantization for the L1-RSRP measurements.

- No constraint on UCI payload overhead for full report of the L1-RSRP measurements of Set B for NW-side models are assumed.

- Observations are applicable for both Tx beam and beam pair.

- The evaluation results are from BM-Case 1 and similar observation are expected for BM-Case 1 when Set B is different from Set A.

Note that, in the following evaluation, model switching is not evaluated for generalization performance.

Companies have provided evaluation results which show that Case 3 and/or Case 2A can provide better performance than Case 2. In most of the cases/evaluations, Case 3 has performance degradation than Case 1. From the evaluation results from some companies and for some scenarios, Case 3 may have similar or slightly higher performance than Case 1:

- 2 sources: for various UE distribution with same or double training data size,

- 1 source: for different ISDs with triple training data size.

**(A) For some cases**, Case 2 **have some performance degradation** than Case 1 in most of the cases/evaluations. In Case 2, AI/ML still can provide better performance than non-AI baseline option 2 (based on the measurements from Set B of beams):

- For various deployment scenarios: UMa/UMi (with the assumption of same down tilt, same or different NLOS probability, same or different ISD, same or different antenna height)

- (Case 2) For generalization Case 2 compared to Case 1,

- With the assumption of same ISD, antenna height and same NLOS probability for UMa/UMi, evaluation results from 4 sources show less than 5% degradation, evaluation results from 4 sources show 5%~10% degradation

- wherein 1 source assumed different UE distribution with same ISD, antenna height, its results show 5%~17% and less than 5% degradation for 100% outdoor UE and 80%/20% in/outdoor UE, respectively, for different combinations of Set B and Set A (i.e., different ratio of Set B/Set A and Set B could be either subset of Set A or different from Set A) for Top-1 beam prediction accuracy, for DL Tx beam prediction.

- With the assumption of different antenna height for UMa/UMi,

- evaluation results from 1 source show about 13% degradation for Top-1 beam prediction accuracy, for DL Tx beam prediction with same ISD

- evaluation results from 1 source show 16%, and 18% degradation for Top-1 beam prediction accuracy, for DL Tx beam and beam pair prediction respectively, with different ISD

- evaluation results from 1 source show about 13% degradation for Top-1 beam prediction accuracy, for both DL Tx beam and beam pair prediction with same ISD, different antenna heights and NLOS probabilities

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 5 sources show less than 5% degradation, and the evaluation results from 1 source show 8% degradation for Top-1 beam prediction accuracy, for DL Tx beam and/or beam pair prediction.

- wherein 1 source assumed different ISD and antenna height and the results show about 8% degradation for Top-1 beam prediction accuracy for both DL Tx beam and beam pair prediction.

- Various deployment scenarios: ISD 200m/ISD 500m

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 3 sources show about 1%~2% degradation, evaluation results from 2 sources show ~9% degradation for Top-1 beam prediction accuracy for DL Tx beam and/or beam pair prediction.

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 1 source show slightly better (1%~2% for Top-1 beam prediction accuracy) performance compared to Case 1 with triple size of training data for DL Tx beam prediction, and, the evaluation results from 1 source show about 1% degradation on Top-1 beam prediction accuracy for beam pair prediction with the same size of training data.

- Various deployment scenarios: 100% outdoor/20%outdoor

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 4 sources show less than 5% degradation, evaluation results from 3 sources show 5%~10% degradation, evaluation results from 3 sources show 10%~25% degradation for Top-1 beam prediction accuracy for DL Tx beam and/or beam pair prediction.

- In addition, 1 source evaluated the scenario with 80% outdoor/20% outdoor, and its evaluation results show about 20% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- In addition, 1 source evaluated the scenario with 100% outdoor/0% outdoor, and its evaluation results show 10%~25% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- In addition, evaluation results from 1 source show that the performance degradation becomes larger with smaller ratio of Set B/Set A.

- wherein, 1 source evaluated the scenario with ISD=200 in UMa for different combinations of Set B and Set A (i.e., different ratio of Set B/Set A and Set B could be either subset of Set A or different from Set A) and the results show 10%~17% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- (Case 2A) For generalization Case 2A compared to Case 1, evaluation results from 1 source show 1%~6% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- wherein, 1 source evaluated the scenario ISD=200 in UMa for different number of epochs and number of data used for finetuning and the results show 1%~6% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- In addition, 1 source evaluated the scenario with 80% outdoor/20% outdoor, and its evaluation results show 3%~8% degradation for Top-1 beam prediction accuracy for DL Tx beam prediction.

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 4 sources show less than 2% degradation, and the evaluation results from 2 sources show 10% degradation for Top-1 beam prediction accuracy compared to Case 1. However, the evaluation results from 1 source show slightly better (about 1% for Top-1 beam prediction accuracy) performance compared to Case 1 with double size of training data.

- In additional, 1 source evaluated the scenario with 80% outdoor/20% outdoor, and its evaluation results show slightly better (about 4% for Top-1 beam prediction accuracy) performance compared to Case 1 with same training data size for DL Tx beam prediction.

- In additional, the evaluation results from 1 source show that for generalization from 100% outdoor to 20% outdoor, 7% degradation for Top-1 beam prediction accuracy compared to Case 1. For generalization from 20% outdoor to 100% outdoor, about 4% degradation for Top-1 beam prediction accuracy compared to Case 1.

- For DL Tx beam prediction only, various UE parameters: different UE codebooks, and/or different UE antenna array dimensions

- (Case 2) For generalization Case 2 compared to Case 1, for Top-1 beam prediction accuracy

- evaluation results from 2 sources show less than 1% performance with different UE codebooks.

- evaluation results from 1 source show about 4% degradation, with different UE codebook, different number of Rx elements and panel location.

- evaluation results from 1 source show about 10% degradation with both different number of UE Rx beams, different number of Rx elements, and about 5% degradation with both different number of UE Rx beams (where in Configuration #A, UE Rx beams are subset of UE Rx beams in Configuration #B), and same number of Rx elements,

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 1 source show 1~2.5% degradation with different number of UE Rx beams, different number of Rx elements and panel location, and evaluation results from 1 source show about 7.5% degradation with both different number of UE Rx beams, different number of Rx elements, for Top-1 beam prediction accuracy.

- For beam pair prediction only, various UE parameters: different number of beams in a seen UE codebook when inference using a subset of Rx beams of training

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 2 sources show 2%~15% degradation Top-1 beam prediction accuracy

- wherein, evaluation results from 1 source show 2% with different number of beams in a seen UE codebook for Top-1 beam prediction accuracy based on the assumption that training by 8 Rx beam and inference by 4 of 8 Rx beam.

- wherein, evaluation results from 1 source show 15% degradation with different number of beams in a seen UE codebook for Top-1 beam prediction accuracy based on the assumption that training by 4 Rx beam and inference by 2 of 4 Rx beam.

**(B) For some cases,** Case 2 have **significant performance degradation** than Case 1 in most of the cases/ evaluations. In Case 2, AI/ML can provide comparable or worse performance than non-AI baseline option 2 (based on the measurements from Set B of beams)

- Various deployment scenarios: UMa/UMi (With the assumption of different ISD, antenna height, down tilt and NLOS probability)

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 3 sources show 20%~35% degradation for Top-1 beam prediction accuracy compared to Case 1, for DL Tx beam and/or beam pair prediction.

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 2 sources show less than 5% degradation,

- Various configurations (parameters and settings): different gNB antenna array dimensions, and/or DL Tx beam codebook

- Note: different DL Tx beam codebooks will result in various Set A of beam(pairs)

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 2 source show 15%~40% degradation, evaluation results from 5 sources show 30%~50% degradation, evaluation results from 2 sources show about 60% degradation, evaluation results from 1 source show about 70% degradation, for Top-1 beam prediction accuracy for DL Tx beam and/or beam pair prediction. 1 source shows BM-AI can perform worse than the conventional approach’s with mismatched set A design.

- Wherein 1 source show 15%-40% degradation for Top-1 beam accuracy assuming same DL Tx codebook (pointing angles) and different beam width, and 50%-60% degradation for Top-1 beam accuracy assuming different DL Tx codebooks (pointing angles) and same beam width for Tx beam and pair prediction

- wherein 2 sources assumed different Tx beam codebooks have different horizontal angles but the same gNB array/beamwidth and the results show about 56% degradation for Top-1 beam prediction accuracy with same training data size for DL Tx beam prediction.

- wherein 1 source assumed different Tx beam codebooks have different horizonal beam angles and the different gNB array/beamwidth and the results show about 57% degradation for Top-1 beam prediction accuracy with same training data size for beam pair prediction.

- wherein 2 sources assumed different Tx beam codebooks have the same beam pointing angles but have different beamwidth (due to different gNB array sizes) and the results show about 30% degradation for Top-1 beam prediction accuracy.

- evaluation results from 1 source show performance degradation in terms of the top-1 beam accuracy from 73.9% to 34.2% at 4 beams in Set B, from 88.6% to 63.9% at 8 beams in set B, from 97.8% to 88.4% at 16 beams in set B.

- evaluation results from 5 sources show better performance than non-AI baseline option 2 (based on the measurements from Set B of beams). However, evaluation results from 5 sources similar or even worse performance than non-AI baseline option 2 (based on the measurements from Set B of beams).

- (Case 2A) For generalization Case 2A compared to Case 1, evaluation results from 1 source show 16%~20% for Top-1 beam prediction accuracy for DL Tx beam prediction with the assumption that different Tx beam codebooks have different horizontal angles but the same gNB array/beamwidth.

- (Case 3) For generalization Case 3 compared to Case 1, the evaluation results from 6 sources show less than 5% degradation, and the evaluation results from 2 sources show 10%~15% degradation for Top-1 beam prediction accuracy compared to Case 1. Evaluation results from 1 source show there is 2%~32% degradation for Top-1 beam with 1 dB margin.

- Wherein, 1 source assumes different beamwidth and double training data size

- For Tx-Rx beam pair prediction only, various UE parameters: different UE codebooks, and/or different UE antenna array dimensions

- Note: different UE Rx beam codebooks will result in various Set A of beam pairs for beam pair prediction

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 4 sources show large degradation (i.e., >40%) with different number of elements (different beamwidth) and different UE codebooks for Top-1 beam prediction accuracy.

- wherein, evaluation results from 1 source show 12% and 52% degradation with UE codebook is different for Top-1 beam prediction accuracy with 1x4 Rx beam and with 2x2 Rx beam pattern and 1x4 Rx beam respectively.

- (Case 3) For generalization Case 3 compared to Case 1, evaluation results from 1 source show less than 5% degradation, and evaluation results from 1 source show 16%~26% degradation for Top-1 beam prediction accuracy, with different number of elements and/or different number of UE Rx

- Various Set B of beams: different fixed Set B pattern

- (Case 2) For generalization Case 2 compared to Case 1, evaluation results from 9 sources show large degradation with different Set B pattern (different number and/or same number different Set B pattern) for DL Tx beam prediction and/or beam pair prediction.

- evaluation results from 1 source show 13~21% degradation with same evenly spaced in beam(pair) ID dimension without providing beam ID information as AI/ML inputs.

- evaluation results from 1 source show 20%~40% degradation with different number of beams in Set B for BMCase-2

- evaluation results from 1 source show the AI-BM performance can be worse than the conventional approach’s with mismatched set B design.

- (Case 3) For generalization Case 3 compared to Case 1,

- evaluation results from 5 sources show less than or about 5% degradation.

- evaluation results from 1 source show 14% degradation without providing beam ID information as AI/ML inputs.

- evaluation results from 1 source show 3%~10% degradation with different number of beams in Set B for BMCase-2

- evaluation results from 1 source show 8-10% degradation with different Set B pattern.

**(C) For BMCase-2, various UE mobility,** different companies reported different observation for Case 2. In Case 2, AI/ML still can provide comparable or worse performance than non-AI baseline option 2 (based on the measurements from Set B of beams)]

- For various UE mobility for BMCase-2: 30km/h / 60km/h / 90km/h 120km/h

- (Case 2) For generalization Case 2 compared to Case 1,

- evaluation results from 3 sources show significant degradation i.e., >30% in terms of Top 1 prediction accuracy, and evaluation results from 1 source show about 19%~49% degradation for prediction time 160ms~800ms.

- evaluation results from 4 sources show >6% performance degradation in terms of Top 1 prediction accuracy and evaluation results from 3 sources show about 10~18% degradation

- (Case 3) For generalization Case 3 compared to Case 1, for Top-1 beam prediction accuracy

- the evaluation results from 3 sources show 3~7% degradation for Top-1 beam prediction accuracy

- the evaluation results from 1 source show 8~14% degradation for Top-1 beam prediction accuracy

- the evaluation results from 1 source show <17% degradation for Top-1 beam prediction accuracy by training with same size of training data mixed of 30km/h, 60km/h and 90km/h.

- the evaluation results from 1 source show about 1% degradation for Top-1 beam prediction accuracy for 30km/h and 60km/h, and show about 4%/8% degradation for Top-1 beam prediction accuracy for 30km/h and 90km/h.

- the evaluation results from 1 source show comparable performance for Top-1 beam prediction accuracy for 30km/h and 60km/h

- the evaluation results from 3 sources show slightly better (1%~2% for Top-1 beam prediction accuracy) performance compared to Case 1 with double or triple size of training data for DL Tx beam prediction.

#### 6.3.2.5 Summary of Performance Results for Beam Management

**Summary of evaluations and results for BM-Case1**

For BM-Case1 when Set B is a subset of Set A or when Set B is different than Set A, without UE rotation, AI/ML can achieve good performance with measurements of fixed Set B that is 1/4 or 1/8 of Set A of beam measured with best Rx beam for DL Tx beam prediction, and with measurements of fixed Set B that is 1/4 or 1/8 or 1/16 of Set A for beam pair prediction. In addition, based on the evaluation results from 2 or 3 sources, for BM-Case1 DL Tx beam prediction, with 1/4 or 1/8 measurement/RS overhead, 96%~99% or 85%~98% of UE average throughput and 95%~97% or 70%~84% of UE 5%ile throughput of non-AI baseline option 1 (exhaustive search over Set A beams) can be achieved according to the predicted beam from AI/ML. Note that, ideal measurements are assumed in the evaluations (in clause 6.3.2.1): beam could be measured regardless their SNR, no measurement error, and measurements obtained in a single-time instance (within a channel-coherence time interval), no quantization and no constraint on UCI payload (for NW-side model).

With some realistic consideration (in clause 6.3.2.3):

- Existing quantization granularity of L1-RSRP causes a minor loss in beam prediction accuracy compared to unquantized L1-RSRPs of beams in Set B at least for BM-Case1 for inference of DL Tx beam prediction.

- Measurement errors degrade the beam prediction performance with AI/ML, while measurement errors also degrade the performance with non-AI baseline (both option 1 and option 2).

- For DL Tx beam prediction, with the measurements from quasi-optimal Rx beam, some performance degradation (e.g., 2% to up to12% Top-1 beam prediction accuracy loss based on most of results) is observed comparing to with measurements from best Rx beam. If the measurements are from random Rx beam, large performance degradation is observed.

In addition, comparing with fixed Set B (Opt 1), in case of with Set B changed among pre-configured patterns (Opt 2B), some performance degradation (e.g., no more than or about 10% Top-1 beam prediction accuracy loss based on most of results) is observed; in case of with Set B randomly changed in Set A of beams (Opt 2C), large degradation (e.g, 20%~50% Top-1 beam prediction accuracy loss based on most of results) is observed. With reduced number of measurements of a fixed set of beams (Set C) as inputs of AI/ML (Opt 2D), some performance degradation (e.g., <10% Top-1 beam prediction accuracy loss based on most of results) is observed, comparing with using all measurements from Set C, in the meanwhile, UCI reporting overhead for inference inputs can be reduced (e.g., 1/2 to 7/8 UCI reporting overhead reduction) comparing with reporting all measurements of the fixed beam Set C.

Moreover, the performance with different label options has been evaluated which may lead to different data collection overhead for training (for both BM-Case1 and BM-Case2).

**Summary of evaluations and result for BM-Case2**

Evaluation results for BM-Case2 when Set B= Set A for DL Tx beam prediction with the measurements from the best Rx beam and beam pair prediction are summarized in Table 6.3.2.5-1 and Table 6.3.2.5-2, without considering generalization aspects.

Table 6.3.2.5-1: Summary of the evaluation results for BM-Case2
when Set B=Set A for DL Tx beam prediction

|  |  |  |
| --- | --- | --- |
|  | Without rotation | With rotation |
| Beam prediction accuracy performance compared with non-AI baseline (option 2) | For 80ms or 160ms prediction time:- Some evaluation results show AI/ML may have similar performance or some degradationFor 160ms or larger prediction time:- Most evaluation results show AI/ML provides some beam prediction accuracy gain- The longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML | AI/ML can provide some beam prediction accuracy gain:- The longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML(2 sources) |
| RS overhead Case A, compared with non-AI baseline (option 1) | AI/ML can achieve decent beam prediction accuracy with 1/5~1/2 measurement/RS overhead reduction | N/A |
| RS overhead Case B, comparing with non-AI baseline (option 2) with given prediction accuracy | AI/ML can achieve a certain beam prediction accuracy with 7/10 measurement/RS overhead reduction(1 source) | AI/ML can achieve a certain beam prediction accuracy with 1/2 measurement/RS overhead reduction(1 source) |
| RS overhead Case B+, comparing with non-AI baseline (option 1) | AI/ML can achieve good beam prediction with 80% measurement/RS overhead reduction(1 source) | AI/ML can achieve good beam prediction with more than 80% measurement/RS overhead reduction(1 source) |

Table 6.3.2.5-2: Summary of the evaluation results for BM-Case2
when Set B=Set A for beam pair prediction

|  |  |  |
| --- | --- | --- |
|  | Without rotation | With rotation |
| Beam prediction accuracy performance compared with non-AI baseline (option 2) | For 160ms or less prediction time:- AI/ML may or may not provide beam prediction accuracy gainThe longer the prediction time,- the higher gain of beam prediction accuracy can be achieved by AI/ML | AI/ML may or may not provide beam prediction accuracy gain(3 sources) |
| RS overhead Case A, compared with non-AI baseline (option 1) | AI/ML can provide good beam prediction accuracy with the less measurements/RS overhead (up to 1/2) | N/A |
| RS overhead Case B, comparing with non-AI baseline (option 2) with given prediction accuracy | AI/ML can achieve a certain beam prediction accuracy with 1/2 or 3/5 measurement/RS overhead reduction(2 sources) | N/A |
| RS overhead Case B+, comparing with non-AI baseline (option 1) | AI/ML can achieve good beam prediction accuracy with 80% measurement/RS overhead reduction(1 source) | N/A |

For BM-Case2 when Set B is a subset of Set A for DL Tx beam prediction with the measurements from the best Rx beam, without considering generalization aspects, AI/ML can achieve good prediction accuracy with 1/2, 1/3, 1/4, 1/8 RS overhead in spatial domain, for the case Set B is fixed or variable with pre-configured patterns of beams with or without UE rotation. More RS/measurements overhead reduction can be achieved considering overhead reduction in time domain.

For BM-Case2 when Set B is a subset of Set A for beam pair prediction, without considering generalization aspects

- without UE rotation, AI/ML can achieve good prediction accuracy with 1/4, 1/8, 1/16 RS overhead in spatial domain, for the case Set B is fixed or variable with pre-configured patterns of beams.

- with UE rotation, from 2 sources, AI/ML can provide 15% or 44% prediction accuracy gain with 1/4, 1/16 RS overhead in spatial domain comparing with non-AI baseline (option 2), for the case Set B is fixed or variable with pre-configured patterns of beams. However, the Top-1 beam prediction accuracy may or may not be good enough.

- More RS/measurements overhead reduction can be achieved considering overhead reduction in time domain.

Note that, ideal measurements are assumed in the above evaluations (for BM-Case2): beam could be measured regardless their SNR, no measurement error, no quantization and no constraint on UCI payload (for NW-side model). With measurement error, quantization or measurements results from quasi-optimal Rx beam for DL Tx beam prediction, similar observations are observed (for some cases) or expected as for BM-Case1.

Reduced measurement overhead can reduce measurement latency for beam prediction in some configurations.

**Summary of evaluations and results for generalization**

Different location of AI/ML model (e.g., NW side model, or UE side model) may have different generalization requirements:

For NW side model,

- generalization performance with various gNB settings and various Set B of beams may not be an issue since the gNB settings are most likely to be fixed or limited to a given gNB (at least seen by AI/ML before)

- for DL Tx beam prediction, generalization performance with various unseen UE parameters is acceptable at least with the measurement from the best or fixed Rx beam.

- Tx-Rx beam pair prediction, generalization performance with various UE parameters, i.e., different number of beams in a seen UE codebook when inference using a subset of Rx beams of training is acceptable~~.~~

- for Tx-Rx beam pair prediction, the significant generalization performance degradation with unseen various UE parameters (i.e., different UE codebooks, and/or different UE antenna array dimensions) can be improved to achieve less than 5% degradation (2 sources) and 16%~26% degradation (1 source) in terms of Top-1 beam prediction accuracy with the model training with mixed data compared to generalization performance Case 1.

- Note: with same amount of data for training for different scenarios for Case 3

- Alternatively, AI/ML model can be trained for different scenarios and rely on model switching based on applicable scenario which would improve generalization performance.

For UE side model,

- generalization performance with unseen various UE parameters may not be an issue

- the significant generalization performance degradation with unseen various gNB setting (i.e., different gNB antenna array dimensions, and/or DL Tx beam codebook) or unseen various Set B of beam(pairs) can be improved to achieve

- (for gNB setting) less than 5% (6 sources), 10%~15% (2 sources), and 2%~32% (1 source) degradation in terms of Top-1 beam prediction accuracy compared with the model training with mixed data to generalization performance Case 1, and 16%~20% (1 source) degradation in terms of Top-1 beam prediction accuracy compared with the model finetune to generalization performance Case 1.

- (for Set B of beam(pairs)) less than 10% (all 7 sources) degradation in terms of Top-1 beam prediction accuracy compared with the model training with mixed data to generalization performance Case 1.

- Note: For gNB setting, generalization performance Case 3 may depend on how different the gNB settings are across training and inference

- Note: with same amount of data for training for different scenarios for Case 3

- Alternatively, AI/ML model can be trained for different scenarios and rely on model switching based on applicable scenario which would improve generalization performance.

At least for BMCase-1, AI/ML (without considering model switching) has some performance degradation with some unseen scenarios including:

- For DL Tx beam prediction,

- deployment scenarios: different ISD, UMi/UMa (at least with same down tilt)

- various outdoor/indoor UE distributions

- various UE parameters: different UE codebooks, and different UE antenna array dimensions.

- Note: at least with the measurement from the best Rx beam.

- For beam pair prediction

- deployment scenarios: different ISD, UMi/UMa (at least with same down tilt)

- various outdoor/indoor UE distributions

- various UE parameters: when inference using a subset of Rx beams of training.

However, the AI/ML (without considering model switching) has significant performance degradation with some other unseen scenarios, including:

- For DL Tx beam prediction,

- deployment scenarios: UMi/UMa (at least with the assumption of different ISD, antenna height, down tilt and NLOS probability)

- various gNB setting: different gNB antenna array dimensions, and DL Tx beam codebook

- various Set B patterns

- various Set A patterns

- For beam pair prediction

- various UE parameters: different UE codebooks, and different UE antenna array dimensions

- deployment scenarios: with the assumption of different ISD, antenna height, down tilt and NLOS probability

- various gNB setting: different gNB antenna array dimensions, and DL Tx beam codebook

- various Set B patterns

- various Set A patterns

In order to let AI/ML model see the data from a new setting which causes performance loss, the AI/ML model can be trained with mixed data or finetuned with the data from the new setting to improve the generalization performance. Alternatively, AI/ML model can be trained for different scenarios and rely on model switching based on applicable scenario which would improve generalization performance.

**For BMCase-2,** for variable UE mobility, the collected data for training can be mixed and the generalization performance with mixed UE speeds is acceptable.

## 6.4 Positioning accuracy enhancements

### 6.4.1 Evaluation assumptions, methodology and KPIs

For AI/ML-based positioning evaluation, RAN1 does not attempt to define any common AI/ML model as a baseline. In the evaluation, some results use UE measurement information as model input, other results use gNB measurement information as model input, and they are not distinguished for summarizing the results.

For AI/ML based positioning, the following methods are evaluated.

(1) Direct AI/ML positioning, see an example illustrated in Figure 6.4.1-1.

(2) Assisted AI/ML positioning.

(a) Assisted AI/ML positioning with multi-TRP construction, see an example illustrated in Figure 6.4.1-2.

(b) Assisted positioning with single-TRP construction and one model for N TRPs, see an example illustrated in Figure 6.4.1-3.

(c) Assisted positioning with single-TRP construction and N models for N TRPs, see an example illustrated in Figure 6.4.1-4.



Figure 6.4.1-1: Direct AI/ML positioning



Figure 6.4.1-2: Assisted positioning with multi-TRP construction



Figure 6.4.1-3: Assisted positioning with single-TRP construction, and one model for N TRPs.



Figure 6.4.1-4: Assisted positioning with single-TRP construction, and N models for N TRPs.

***KPIs*:**

- For all scenarios and use cases, the main KPI is the CDF percentiles of horizonal accuracy

- The CDF percentiles to analyse are: 90% (baseline) and {50%, 67%, 80%} (optional)

- Vertical accuracy can be optionally reported

- Target positioning requirements for horizonal accuracy and vertical accuracy are not defined for AI/ML-based positioning evaluation

- Model complexity, e.g., number of model parameters, and computational complexity, e.g., FLOPs

- Model complexity is reported via the metric of "number of model parameters". Note: if complex value is used in modelling process, the number of the model parameters is doubled, which is also applicable for other AIs of AI/ML.

- Companies describe how their computational complexity values are obtained. It is out of 3GPP scope to consider computational complexity values that have platform-dependency and/or use implementation (hardware and software) optimization solutions.

- For AI/ML assisted positioning, an intermediate performance metric of *model output*

***Model generalization*:**

To investigate the model generalization capability, at least the following aspect(s) are considered for the evaluation for AI/ML based positioning:

- Different drops: Training dataset from drops {A0, A1,…, AN-1}, test dataset from unseen drop(s) (i.e., different drop(s) than any in {A0, A1,…, AN-1}). Here N≥1.

- Clutter parameters, e.g., training dataset from one clutter parameter (e.g., {40%, 2m, 2m}), test dataset from a different clutter parameter (e.g., {60%, 6m, 2m});

- Network synchronization error, e.g., training dataset without network synchronization error, test dataset with network synchronization error;

- UE/gNB RX and TX timing error;

- The baseline non-AI/ML method may enable the Rel-17 enhancement features (e.g., UE Rx TEG, UE RxTx TEG).

- InF scenarios, e.g., training dataset from one InF scenario (e.g., InF-DH), test dataset from a different InF scenario (e.g., InF-HH)

- The following issues related to measurements on the positioning accuracy of the AI/ML model. The simulation assumptions reflecting these issues are up to companies.

- SNR mismatch (i.e., SNR when training data are collected is different from SNR when model inference is performed).

- Time varying changes (e.g., mobility of clutter objects in the environment)

- Channel estimation error

For both direct AI/ML approach and AI/ML assisted approach, for a given AI/ML model design (e.g., input, output, single-TRP vs multi-TRP), identify the generalization aspects where model fine-tuning/mixed training dataset/model switching is necessary.

***Evaluation assumptions*:**

The IIoT indoor factory (InF) scenario is a prioritized scenario for evaluation of AI/ML based positioning. Specifically, InF-DH sub-scenario is prioritized for FR1 and FR2.

Reuse the common scenario parameters defined in Table 6-1 of TR 38.857. For evaluation of InF-DH scenario, the parameters are modified from TR 38.857 Table 6.1-1 as shown in Table 6.4.1-1. The parameters in the table are applicable to InF-DH at least. If other InF sub-scenario is evaluated in addition to InF-DH, some parameters in Table 6-5 may be updated. If an InF scenario different from InF-DH is evaluated for the model generalization capability, the selected parameters (e.g., clutter parameters) are compliant with TR 38.901 Table 7.2-4 (Evaluation parameters for InF). Note: In TR 38.857 Table 6.1-1 (Parameters common to InF scenarios), InF-SH scenario uses the clutter parameter {20%, 2m, 10m} which is compliant with TR 38.901.

Table 6.4.1-1: Parameters common to InF scenario (Modified from TR 38.857 Table 6.1-1) for AI/ML based positioning evaluations

|  |  |  |
| --- | --- | --- |
|  | FR1 specific values | FR2 specific values |
| Channel model | InF-DH | InF-DH |
| Layout | Hall size | InF-DH: (baseline) 120x60 m(optional) 300x150 m |
| BS locations | 18 BSs on a square lattice with spacing D, located D/2 from the walls.- for the small hall (L=120m x W=60m): D=20m- for the big hall (L=300m x W=150m): D=50m |
| Room height | 10 m |
| Total gNB TX power, dBm | 24dBm | 24dBmEIRP should not exceed 58 dBm |
| gNB antenna configuration | (M, N, P, Mg, Ng) = (4, 4, 2, 1, 1), dH=dV=0.5λ according to Table A.2.1-7 in TR 38.802.Note: Other gNB antenna configurations are not precluded for evaluation. | (M, N, P, Mg, Ng) = (4, 8, 2, 1, 1), dH=dV=0.5λ according to Table A.2.1-7 in TR 38.802.One TXRU per polarization per panel is assumed. |
| gNB antenna radiation pattern | Single sector according to Table A.2.1-7 in TR 38.802. | 3-sector antenna configuration according to Table A.2.1-7 in TR 38.802 |
| Penetration loss | 0dB |
| Number of floors | 1 |
| UE horizontal drop procedure | Uniformly distributed over the horizontal evaluation area for obtaining the CDF values for positioning accuracy, The evaluation area should be selected from- (baseline) the whole hall area, and the CDF values for positioning accuracy is obtained from whole hall area.- (optional) the convex hull of the horizontal BS deployment, and the CDF values for positioning accuracy is obtained from the convex hull. |
| UE antenna height | Baseline: 1.5m(Optional): uniformly distributed within [0.5, X2] m, where X2 = 2m for scenario 1 (InF-SH) and X2= *hc* for scenario 2 (InF-DH)  |
| UE mobility | 3km/h  |
| Min gNB-UE distance (2D), m | 0m |
| gNB antenna height | Baseline: 8m(Optional): two fixed heights, either {4, 8} m, or {max(4, *hc*), 8}. |
| Clutter parameters: {density *r*, height *hc*, size *dclutter*} | High clutter density:- {60%, 6m, 2m}- {40%, 2m, 2m} - can be considered optional in the evaluations considering specific AI/ML designs. |
| Channel Estimation | Assumption, e.g., realistic or ideal channel estimation, error models, receiver algorithms should be reported.  |
| Spatial consistency | If enabled for the evaluations:Model at least one of: large scale parameters, small scale parameters and absolute time of arrival, where:- the large scale parameters are according to clause 7.5 of TR 38.901 and correlation distance = *dclutter*/2 for InF (clause 7.6.3.1 of TR 38.901)- the small scale parameters are according to clause 7.6.3.1 of TR 38.901- the absolute time of arrival is according to clause 7.6.9 of TR 38.901Baseline evaluation does not incorporate spatially consistent UT/BS mobility modelling (clause 7.6.3.2 of TR 38.901). It is optional to implement it. |
| Baseline for performance evaluation | Existing Rel-16/Rel-17 positioning methods. Specific existing positioning method (e.g., DL-TDOA, Multi-RTT) used as comparison is to be reported.  |

For the evaluation of AI/ML based positioning, the study of model input due to different number of TRPs include the following approaches. Proponents of each approach are to provide analysis for model performance, signalling overhead (including training data collection and model inference), model complexity and computational complexity.

- Approach 1: Model input size stays constant as NTRP=18. The number of TRPs (N’TRP) that provide measurements to model input varies. When N’TRP < NTRP, the remaining (NTRP - N’TRP) TRPs do not provide measurements to model input, i.e., measurement value is set such that the (NTRP − N’TRP) TRPs do not affect model output.

- Approach 1-A. The set of TRPs (N’TRP) that provide measurements is fixed.

- Approach 1-B. The set of TRPs (N’TRP) that provide measurements can change dynamically.

- Note: for Approach 1, one model is provided to cover the entire evaluation area.

- Approach 2: The TRP dimension of model input is equal to the number of TRPs (N’TRP) that provide measurements as model input. When N’TRP < NTRP, the remaining (NTRP - N’TRP) TRPs are ignored by the given model.

- Approach 2-A. The set of active TRPs (N’TRP) that provide measurements is fixed.

- For both Approach 1-A and 2-A: one model can be provided to cover the entire evaluation area, which is equivalent to deploying N’TRP TRPs in the evaluation area for positioning if ignoring the potential inference from the remaining (18 - N’TRP) TRPs.

- Approach 2-B: The set of active TRPs (N’TRP) that provide measurements can change dynamically.

- For Approach 2-B, one model is developed to handle various patterns of active TRPs.

- For Approach 2, if Nmodel (Nmodel >1) models are provided to cover the entire evaluation area, the total model complexity is the summation of the Nmodel models.

In the evaluation of AI/ML based positioning, if N’TRP<18, the set of N’TRP TRPs that provide measurements to model input of an AI/ML model are reported using the TRP indices shown in Figure 6.4.1-5.



Figure 6.4.1-5: TRP layout and their indices for the evaluation of AI/ML based positioning.

For the evaluation of AI/ML based positioning method, the measurement size and signalling overhead for the model input are reported.

Impact from implementation imperfections is to be studied. Further, how AI/ML positioning accuracy is affected by user density/size of the training dataset is to be also studied. Note: details of user density/size of training dataset to be reported in the evaluation.

***Model input, model output:***

For the model input used in evaluations of AI/ML based positioning, if time-domain channel impulse response (CIR) or power delay profile (PDP) is used as model input in the evaluation, companies report the input dimension NTRP \* Nport \* Nt, where NTRP is the number of TRPs, Nport is the number of transmit/receive antenna port pairs, Nt is the number of consecutive time domain samples. If N’t (N’t < Nt) samples with the strongest power are selected as model input, with remaining (Nt ‒ N’t) time domain samples set to zero, then companies report value N’t in addition to Nt. It is also assumed that timing info for the N’t samples need to be provided as model input.

For evaluation of AI/ML based positioning, when time domain samples are used as model input and sub-sampling is applied, the selection of N't measurements is based on the strongest power, unless explicitly stated otherwise. When sub-sampling is applied the N't measurement are not necessarily consecutive in time.

- Training dataset and test dataset use the same measurement selection method (e.g., strongest power) unless explicitly stated otherwise.

- Other selection methodologies for N't measurements are also evaluated, and are not precluded.

For evaluations, companies used the following values for sampling period:

- 16 Sources used the following sampling period:

- Sampling period = 1/(Nf ×∆f). For FR1, sampling period = 1/(4096×30)=8.14 (ns), where Nf =4096 according to 38.211, and ∆f =30 kHz is the subcarrier spacing.

- 1 Source used: sampling period = 4.069 ns

If the model input is the CIR, then each input value of the CIR is a complex number, i.e., it contains two real values, either {real, imaginary} or {magnitude, phase}. If the model input is the PDP, then each input value of the PDP is a real value. Optionally companies can use delay profile (DP) as a type of information for model input. DP is a degenerated version of PDP, where the path power is not provided.

Note: CIR and PDP may have different dimensions. Companies to provide details on their assumption on how PDP is constructed and how (if applicable) it is mapped to Nt samples.

For evaluation of AI/ML based positioning, when timing information is included in model input (e.g., in CIR/PDP/DP), training dataset and test dataset use the same timing format (i.e., both are absolute time, or both are relative time) unless explicitly stated otherwise.

For evaluation of AI/ML based positioning with multipath measurement for model input,

- For a given set of parameters (N'TRP, Nt, N't, Nport)

- CIR has the largest measurement size, where CIR is composed of a list of measurements where each measurement contains the information of: (a) delay, (b) power and (c) phase.

- PDP has smaller measurement size than CIR, where PDP is composed of a list of measurements where each measurement contains the information of: (a) delay and (b) power.

- DP has the smallest measurement size, where DP is composed of a list of measurements where each measurement contains the information of: (a) delay.

- For each model input type (CIR, PDP, DP)

- The measurement size increases (approximately) linearly as N'TRP increases, where N'TRP is the number of active TRPs that provide measurements for the positioning.

- The measurement size increases (approximately) linearly as Nport increases, where Nport is the number of transmit/receive antenna port pairs that provide measurements for the positioning.

- If N't (N't < Nt) measurements are selected as model input, measurement size for model input increases (approximately) linearly with N't;

- For model input type CIR and PDP, if the full set of Nt measurements in time domain is used as model input, measurement size for model input increases (approximately) linearly with Nt;

- Note: if DP is used as model input, DP does not use full set of Nt measurements in time domain (i.e., N't < Nt always).

- Note: for Case 2b and 3b, measurement size of model input has impact to signalling overhead for model inference, data collection, and monitoring.

- Note: There are trade-offs between measurement size / signalling overhead and positioning accuracy when using different sets of parameters (N'TRP, Nt, N't, Nport).

For both the direct AI/ML positioning and AI/ML assisted positioning, the model input is studied, considering the trade-off among model performance, model complexity and computational complexity:

- The type of information to use as model input. The candidates include at least: time-domain CIR, PDP.

- The dimension of model input in terms of NTRP, Nt, and Nt’.

- Note: For the direct AI/ML positioning, model input size has impact to signalling overhead for model inference

At least for model inference of AI/ML assisted positioning, evaluate and report the AI/ML model output, including:

a) the type of information (e.g., ToA, RSTD, AoD, AoA, LOS/NLOS indicator) to use as model output,

b) soft information vs hard information,

c) whether the model output can reuse existing measurement report (e.g., NRPPa, LPP).

***Labels:***

The performance impact from availability of the ground-truth labels (i.e., some training data may not have ground-truth labels) is to be studied. The learning algorithm (e.g., supervised learning, semi-supervised learning) is to be reported by participating companies and, when providing evaluation results, data labelling details need to be described, including:

- Meaning of the label (e.g., UE coordinates; binary identifier of LOS/NLOS; ToA)

- Percentage of training data without label, if incomplete labelling is considered in the evaluation

- Imperfection of the ground-truth labels, if any

Whether, and if so how, an entity can be used to obtain ground-truth label and/or other training data is to be studied.

For direct AI/ML positioning, the impact of labelling error to positioning accuracy is studied considering:

- The ground-truth label error in each dimension of x-axis and y-axis can be modelled as a truncated Gaussian distribution with zero mean and standard deviation of L meters, with truncation of the distribution to the [-2\*L, 2\*L] range. Value L is up to sources.

For AI/ML assisted positioning with TOA as model output, study the impact of labelling error to TOA accuracy and/or positioning accuracy.

- The ground-truth label error of TOA is calculated based on location error. The location error in each dimension of x-axis and y-axis can be modelled as a truncated Gaussian distribution with zero mean and standard deviation of L meters, with truncation of the distribution to the [-2\*L, 2\*L] range.

- Value L is up to sources.

- Other models of labelling error are not precluded

- Other timing information, e.g., RSTD, as model output is not precluded.

For AI/ML assisted positioning with LOS/NLOS indicator as model output, study the impact of labelling error to LOS/NLOS indicator accuracy and/or positioning accuracy.

- The ground-truth label error of LOS/NLOS indicator can be modelled as m% LOS label error and n% NLOS label error. Value m and n are up to sources.

- m%=FN/NLOS is false negative rate of the training data label, where FN (False Negative) is the number of actual LOS links which are incorrectly labelled as NLOS, and NLOS is the total number of actual LOS links;

- n%=FP/NNLOS is the false positive rate of the training data label, FP (False Positive) is the number of actual NLOS links which are incorrectly labelled as LOS, and NNLOS is the total number of actual NLOS links

- Companies consider at least hard-value LOS/NLOS indicator as model output.

***Training dataset:***

Synthetic dataset generated according to the statistical channel models in TR 38.901 is used for model training, validation, and testing. The dataset is generated by a system level simulator based on 3GPP simulation methodology.

As a starting point, the training, validation and testing dataset are from the same large-scale and small-scale propagation parameters setting. Subsequent evaluations can study the performance when the training dataset and testing dataset are from different settings.

Details of the training dataset generation are to be reported, including:

- The size of training dataset, e.g., the total number of UEs in the evaluation area for generating training dataset;

- The distribution of UE location for generating the training dataset may be one of the following:

- Option 1: grid distribution, i.e., one training data is collected at the center of one small square grid, where, for example, the width of the square grid can be 0.25/0.5/1.0 m.

- Option 2: uniform distribution, i.e., the UE location is randomly and uniformly distributed in the evaluation area.

***Sub-use case specific*:**

For AI/ML-assisted positioning, companies report which construction is applied in their evaluation:

a) Single-TRP construction: the input of the ML model is the channel measurement between the target UE and a single TRP, and the output of the ML model is for the same pair of UE and TRP.

b) Multi-TRP construction: the input of the ML model contains N sets of channel measurements between the target UE and N (N>1) TRPs, and the output of the ML model contains N sets of values, one for each of the N TRPs.

Notes: For a measurement (e.g., RSTD) which is a relative value between a given TRP and a reference TRP, the TRP in "single-TRP" and "multi-TRP" refers to the given TRP only. For single-TRP construction, companies report whether they consider same model for all TRPs or N different models for TRPs.

When single-TRP construction is used for the AI/ML model, companies report at least the AI/ML complexity (Model complexity, Computation complexity) for N TRPs, which are used to determine the position of a target UE considering the various constructions in Table 6.4.1-2 below.

Table 6.4.1-2: Model complexity and computational complexity to support N TRPs for a target UE

|  |  |  |
| --- | --- | --- |
|  | Model complexity to support N TRPs | Computational complexity to process N TRPs |
| Single-TRP, same model for N TRPs | $$P\_{S}$$where $P\_{S}$ is the model complexity for one TRP and the same model is used for N TRPs. | $$N×C\_{S}$$where $C\_{S}$ is the computation complexity of the same model for one TRP. |
| Single-TRP, N models for N TRPs | $$\sum\_{i=1,…N}^{}P\_{S,i}$$where $P\_{S,i}$ is the model complexity for the i-th AI/ML model. | $$\sum\_{i=1,…N}^{}C\_{S,i}$$where $C\_{S,i}$ is the computation complexity for the i-th AI/ML model. |
| Multi-TRP (i.e., one model for N TRPs) | $$P\_{M}$$where $P\_{M}$ is the model complexity for the one model. | $$C\_{M}$$where $C\_{M}$ is the computation complexity for the one model. |

Note: The reported model complexity above is intended for inference and may not be directly applicable to complexity of other LCM aspects

For evaluation of AI/ML assisted positioning, the following intermediate performance metrics are used:

- LOS classification accuracy, if the model output includes LOS/NLOS indicator of hard values, where the LOS/NLOS indicator is generated for a link between UE and TRP;

- Timing estimation accuracy (expressed in meters), if the model output includes timing estimation (e.g., ToA, RSTD).

- Angle estimation accuracy (in degrees), if the model output includes angle estimation (e.g., AoA, AoD).

- Companies provide info on how LOS classification accuracy and timing/angle estimation accuracy are estimated, if the ML output is a soft value that represents a probability distribution (e.g., probability of LOS, probability of timing, probability of angle, mean and variance of timing/angle, etc.)

***Model monitoring:***

For AI/ML assisted approach, the performance of model monitoring metrics is studied at least where the metrics are obtained from inference accuracy of model output (i.e., label-based model monitoring methods). Further, the performance of label-free model monitoring methods, which do not require ground-truth label (or its approximation) for model monitoring, is to be studied.

For direct AI/ML positioning, the performance of model monitoring methods is studied, including:

- Label based methods, where ground-truth label (or its approximation) is provided for monitoring the accuracy of model output.

- Label-free methods, where model monitoring does not require ground-truth label (or its approximation).

***Model Fine-tuning***:

For evaluation of the potential performance benefits of model finetuning, training dataset setting (e.g., training dataset size necessary for performing model finetuning) and horizontal positioning accuracy (in meters) before and after model finetuning, are to be reported.

For both direct and AI/ML assisted positioning methods, investigate at least the impact of the amount of fine-tuning data on the positioning accuracy of the fine-tuned model. The fine-tuning data is the training dataset from the target deployment scenario.

### 6.4.2 Performance results

POS\_Table 1 through POS\_Table 5 in attached Spreadsheets for Positioning accuracy enhancements evaluations present the performance results for:

- POS\_Table 1. Evaluation results for supervised learning without generalization considerations (i.e., same setting for training and testing).

- POS\_Table 2. Evaluation results for supervised learning with generalization considerations (i.e., different setting for training and testing).

- POS\_Table 3. Evaluation results for fine-tuning to handle various generalization aspects

- POS\_Table 4. Evaluation results for supervised learning with label error

- POS\_Table 5. Evaluation results for semi-supervised learning

***Observations*:**

Direct AI/ML positioning can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods when the generalization aspects are not considered.

For InF-DH with clutter parameter setting {60%, 6m, 2m}, evaluation results indicate that the direct AI/ML positioning can achieve horizontal positioning accuracy of <1m at CDF=90%, as compared to >15m for conventional positioning methods.

Based on evaluation results of 3 sources, direct AI/ML positioning and AI/ML assisted positioning can achieve comparable performance when simulation assumptions and parameters (e.g., clutter parameter, model input type, model input size, training dataset size, model complexity) are held the same, *E*direct = (0.57~1.14) × *E*assisted, where

*- E*assisted (meters) is the horizontal positioning accuracy at CDF=90% of AI/ML assisted positioning with multi-TRP construction with timing information as model output,

*- E*direct (meters) is the horizontal positioning accuracy at CDF=90% of direct AI/ML positioning

AI/ML assisted positioning can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods when the generalization aspects are not considered.

- For InF-DH with clutter parameter setting {40%, 2m, 2m}, evaluation results indicate that the AI/ML assisted positioning can achieve horizontal positioning accuracy of <0.4m at CDF=90%, as compared to >9m for conventional positioning method.

- For InF-DH with clutter parameter setting {60%, 6m, 2m}, evaluation results indicate that the AI/ML assisted positioning can achieve horizontal positioning accuracy of <1m at CDF=90%, as compared to >15m for conventional positioning method.

***Model monitoring***

For AI/ML assisted positioning, evaluation results have been provided by sources for label-based model monitoring methods. With TOA and/or LOS/NLOS indicator as model output, the estimated ground-truth label (i.e., TOA and/or LOS/NLOS indicator) is provided by the location estimation from the associated conventional positioning method. The associated conventional positioning method refers to the method which utilizes the AI/ML model output to determine target UE location.

For both direct AI/ML and AI/ML assisted positioning, evaluation results have been provided by sources to demonstrate the feasibility of label-free model monitoring methods.

***Model complexity and computational complexity***

For AI/ML based positioning method, companies have submitted evaluation results to show that for their evaluated cases, for a given company’s model design, a lower complexity (model complexity and computational complexity) model can still achieve acceptable positioning accuracy (e.g., <1m), albeit degraded, when compared to a model with higher AI/ML complexity.

In Figure 6.4.2-1 below, the model inference complexity reported by companies for the positioning use case is shown, including (a) on the x-axis: model complexity in number of real parameters (millions) and (b) on the y-axis: computational complexity in FLOPs (millions). Figure 6.4.2-1 shows the range of complexity for the following schemes: (1) direct positioning; (2) assisted positioning with multi-TRP; (3) assisted positioning with single-TRP and one-model for N TRPs; and (4) assisted positioning with single-TRP and N models for N TRPs. For details of the complexity values corresponding to Figure 6.4.2-1, please see POS\_Table 1.

For the three schemes of AI/ML assisted positioning, the complexity is calculated according to Table 6.4.1-2. Both model complexity and computational complexity values are as reported by participating companies. There is no effort to align the procedure across companies on how the complexity values are obtained. In addition, optimizing AI/ML complexity (i.e., model complexity and computational complexity) is out of scope of the study item.



Figure 6.4.2-1: Model complexity and computational complexity for four schemes of AI/ML based positioning.

#### 6.4.2.1 Training Data Collection

***Observations*:**

For data collection of training dataset for AI/ML based positioning, for a given deployment scenario (e.g., InF-scenario, clutter parameter, drop) and with uniform UE distribution, the required sample density (e.g., #samples/m2) for achieving a given positioning accuracy target varies with AI/ML design choices including:

- different positioning approach (direct AI/ML, AI/ML-assisted),

- different type of model input,

- the size of model input,

- AI/ML complexity (model complexity and computational complexity).

For AI/ML based positioning, the positioning accuracy is affected by the training dataset size for a given UE distribution area (or equivalently, sample density in #samples/m2), when the UE is distributed uniformly in training data collection.

- There exists a tradeoff between the training dataset size and the achievable positioning accuracy. The larger the training dataset size (i.e., higher sample density), the smaller the positioning error (in meters), until a saturation point is reached where additional training data does not bring further improvement to the positioning accuracy.

- Note: here a sample refers to the training data collected of one UE at one location. Sample density is equivalent to the density of UEs with data collected in the training dataset.

Evaluation results demonstrate that the performance of AI/ML positioning with the evaluation area as the convex hull of the horizontal BS deployment shows better performance than that with the whole hall area as evaluation area. This is due to: (a) convex hull case has higher sample density if using the same training dataset size, since convex hull has smaller UE distribution area; (b) for whole hall area, the UEs located outside the convex hull have diminished access to TRPs.

- For convex hull: UE distribution area = 100x40 m;

- For whole hall area: UE distribution area = 120x60 m

#### 6.4.2.2 Generalization Aspects

***Observations*:**

***Direct AI/ML positioning***

Evaluation of the following *generalization aspects* show that the positioning accuracy of direct AI/ML positioning deteriorates when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.

- The generalization aspects include:

- Different drops

- Different clutter parameters

- Different InF scenarios

- Network synchronization error

- Companies have provided evaluation results which show that the positioning accuracy on the test dataset can be improved by better training dataset construction and/or model fine-tuning/re-training.

- Better training dataset construction: The training dataset is composed of data from multiple deployment scenarios, which include data from the same deployment scenario as the test dataset.

- Model fine-tuning/re-training: the model is re-trained/fine-tuned with a dataset from the same deployment scenario as the test dataset.

Note: Ideal model training and switching may provide the upper bound of achievable performance when the AI/ML model needs to handle different deployment scenarios.

For direct AI/ML positioning, based on evaluation results of *timing error* in the range of 0-50 ns, when the model is trained by a dataset with UE/gNB RX and TX timing error t1 (ns) and tested in a deployment scenario with UE/gNB RX and TX timing error t2 (ns), for a given t1,

- For a case evaluated by a given source, the positioning accuracy of cases with t2 smaller than t1 is better than the cases with t2 equal to t1. For example,

- For the case of (t1, t2)=(50ns, 30ns), evaluation results show the positioning error of (t1, t2)=(50ns, 30ns) is 0.82~0.86 times that of (t1, t2)=(50ns, 50ns).

- For the case of (t1, t2)=(50ns, 0ns), evaluation results show the positioning error of (t1, t2)=(50ns, 0ns) is 0.80~0.82 times that of (t1, t2)=(50ns, 50ns).

- For a case evaluated by a given source, the positioning accuracy of cases with t2 greater than t1 is worse than the cases with t2 equal to t1. The larger the difference between t1 and t2, the more the degradation. For example,

- For the case of (t1, t2)=(0ns, 10ns), evaluation results show the positioning error of (t1, t2)=(0ns, 10ns) is 1.25~18.7 times that of (t1, t2)=(0ns, 0ns).

- For the case of (t1, t2)=(0ns, 50ns), evaluation results show the positioning error of (t1, t2)=(0ns, 50ns) is 3.5~18.3 times that of (t1, t2)=(0ns, 0ns).

Note: Here the positioning error is the horizonal positioning error (meters) at CDF=90%.

For direct AI/ML positioning, based on evaluation results of *network synchronization* error in the range of 0-50 ns, when the model is trained by a dataset with network synchronization error t1 (ns) and tested in a deployment scenario with network synchronization error t2 (ns), for a given t1,

- For a case evaluated by a given source, the positioning accuracy of cases with t2 smaller than t1 is better than the cases with t2 equal to t1. For example,

- For the case of (t1, t2)=(50ns, 10ns), evaluation results show the positioning error of (t1, t2)=(50ns, 10ns) is 0.52~0.83 times that of (t1, t2)=(50ns, 50ns).

- For the case of (t1, t2)=(50ns, 0ns), evaluation results show the positioning error of (t1, t2)=(50ns, 0ns) is 0.50~0.82 times that of (t1, t2)=(50ns, 50ns).

- For a case evaluated by a given source, the positioning accuracy of cases with t2 greater than t1 is worse than the cases with t2 equal to t1. The larger the difference between t1 and t2, the more the degradation. For example,

- For the case of (t1, t2)=(0ns, 10ns), evaluation results show the positioning error of (0ns, 10ns) is 1.17~9.5 times that of (0ns, 0ns).

- For the case of (t1, t2)=(0ns, 50ns), evaluation results show the positioning error of (0ns, 50ns) is 10~40 times that of (0ns, 0ns).

Note: here the positioning error is the horizonal positioning error (meters) at CDF=90%.

***AI/ML assisted positioning***

For AI/ML assisted positioning with timing information (e.g., ToA) as model output, evaluation of the following *generalization aspects* show that:

- the positioning accuracy deteriorates when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.

- Different drops

- Different clutter parameters

- Different InF scenarios

- the positioning accuracy may or may not deteriorate when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.

- Network synchronization error

- UE/gNB RX and TX timing error

- SNR mismatch

- Channel estimation error

For AI/ML assisted positioning, evaluation results demonstrate that for the *generalization aspects* of:

- Different drops

- Different clutter parameters

- Different InF scenarios

- Network synchronization error

- UE/gNB RX and TX timing error

- SNR mismatch

- Channel estimation error

if the positioning accuracy **would deteriorate** when the AI/ML model is trained with dataset of one deployment scenario and tested with dataset of a different deployment scenario, the positioning accuracy on the test dataset can be improved by better training dataset construction and/or model fine-tuning/re-training.

- Better training dataset construction: The training dataset is composed of data from multiple deployment scenarios, which include data from the same deployment scenario as the test dataset.

- Model fine-tuning/re-training: the model is re-trained/fine-tuned with a dataset from the same deployment scenario as the test dataset.

Note: Ideal model training and switching may provide the upper bound of achievable performance when the AI/ML model needs to handle different deployment scenarios.

For AI/ML assisted positioning with timing information (e.g., ToA) as model output, based on evaluation results of *network synchronization error* in the range of 0-50 ns, when the model is trained by a dataset with network synchronization error t1 (ns) and tested in a deployment scenario with network synchronization error t2 (ns), for a given t1,

- For a case evaluated by a given source, the positioning accuracy of cases with t2 smaller than t1 is better than the cases with t2 equal to t1. For example,

- For the case of (t1, t2)=(50ns, 20~25ns), evaluation results show the positioning error of (t1, t2)=(50ns, 20~25ns) is 0.64~0.85 times that of (t1, t2)=(50ns, 50ns).

- For the case of (t1, t2)=(50ns, 0ns), evaluation results show the positioning error of (t1, t2)=(50ns, 0ns) is 0.50~0.80 times that of (t1, t2)=(50ns, 50ns).

- For a case evaluated by a given source, the positioning accuracy of cases with t2 greater than t1 is worse than the cases with t2 equal to t1. The larger the difference between t1 and t2, the more the degradation. For example,

- For the case of (t1, t2)=(0ns, 10ns), evaluation results show the positioning error of (0ns, 10ns) is 1.16~4.40 times that of (0ns, 0ns).

- For the case of (t1, t2)=(0ns, 20~25ns), evaluation results show the positioning error of (0ns, 50ns) is 2.19~10.11 times that of (0ns, 0ns).

- For the case of (t1, t2)=(0ns, 50ns), evaluation results show the positioning error of (0ns, 50ns) is 9.68~31.95 times that of (0ns, 0ns).

Note: Here the positioning error is the horizonal positioning error (meters) at CDF=90%.

For AI/ML assisted positioning with timing information (e.g., ToA) as model output, based on evaluation results of *timing error* in the range of 0-50 ns, when the model is trained by a dataset with UE/gNB RX and TX timing error t1 (ns) and tested in a deployment scenario with UE/gNB RX and TX timing error t2 (ns), for a given t1,

- For a case evaluated by a given source, the positioning accuracy of cases with t2 smaller than t1 is better than the cases with t2 equal to t1. For example,

- For the case of (t1, t2)=(50ns, 20~25ns), evaluation results ~~submitted to RAN1#113~~ show the positioning error of (t1, t2)=(50ns, 20~25ns) is 0.75~1.00 times that of (t1, t2)=(50ns, 50ns).

- For the case of (t1, t2)=(50ns, 0ns), evaluation results ~~submitted to RAN1#113~~ show the positioning error of (t1, t2)=(50ns, 0ns) is 0.76~0.99 times that of (t1, t2)=(50ns, 50ns).

- For a case evaluated by a given source, the positioning accuracy of cases with t2 greater than t1 is worse than the cases with t2 equal to t1. The larger the difference between t1 and t2, the more the degradation. For example,

- For the case of (t1, t2)=(0ns, 10ns), evaluation results ~~submitted to RAN1#113~~ show the positioning error of (t1, t2)=(0ns, 10ns) is 1.34~5.43 times that of (t1, t2)=(0ns, 0ns).

- For the case of (t1, t2)=(0ns, 20~25ns), evaluation results ~~submitted to RAN1#113~~ show the positioning error of (t1, t2)=(0ns, 20~25ns) is 5.66~13.0 times that of (t1, t2)=(0ns, 0ns).

- For the case of (t1, t2)=(0ns, 50ns), evaluation results ~~submitted to RAN1#113~~ show the positioning error of (t1, t2)=(0ns, 50ns) is 10.62~51.52 times that of (t1, t2)=(0ns, 0ns).

Note: Here the positioning error is the horizonal positioning error (meters) at CDF=90%.

***Both direct AI/ML positioning and AI/ML assisted positioning***

For **both** direct AI/ML and AI/ML assisted positioning, evaluation results submitted show that with CIR model input for a trained model,

- For two SNR/SINR values S1 (dB) and S2 (dB), S1 ≥ S2 + 15 dB, positioning error of a model trained with data of S1 (dB) and tested with data of S2 (dB) is more than 5.75 times that of the model trained and tested with data of S1 (dB).

- For two SNR/SINR values S1 (dB) and S2 (dB), S1 ≤ S2 – 10 dB, the generalization performance of a model trained with data of S1 (dB) and tested with data of S2 (dB) is better than the performance of a model trained with data of S2 (dB) and tested with data of S1 (dB). Positioning error of a model trained with data of S2 (dB) and tested with data of S1 (dB) is more than 2.97 times that of the model trained with data of S1 (dB) and tested with data of S2 (dB).

Note: Here the positioning error is the horizonal positioning error (meters) at CDF=90%.

It is concluded that no dedicated evaluation is needed for the positioning accuracy performance of model switching.

#### 6.4.2.3 Fine-tuning

***Observations*:**

***Direct AI/ML positioning***

For **direct** AI/ML positioning and **different drops**, evaluation has been performed where the AI/ML model is (a) previously trained for drop A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for drop B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under drop B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 6 sources when fine-tuning dataset size is *x*% = 1.3%~2.5% of full training dataset size, the positioning error is *E* = (3.15~10.89) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = 4.0%~5.0% of full training dataset size, the positioning error is *E* = (2.20~8.82) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = 6.3%~10.0% of full training dataset size, the positioning error is *E* = (1.99~7.21) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = 12.0%~25.0% of full training dataset size, the positioning error is *E* = (1.58~5.13) × *E0,B*; 1 source the positioning error is *E* = (10.46) × *E0,B*;

- 3 sources when fine-tuning dataset size is *x*% = 34.0%~50.0% of full training dataset size, the positioning error is *E* = (1.22~2.70) × *E0,B*; 1 source the positioning error is *E* = (8.88) × *E0,B*;

- 2 sources when fine-tuning dataset size is *x*% = 100.0% of full training dataset size, the positioning error is *E* = (1.00~1.19) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for drop B.

For **direct** AI/ML positioning and **different drops**, evaluation has been performed where the AI/ML model is (a) previously trained for drop A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for drop B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under drop A and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 3 sources when fine-tuning dataset size is *x*% = 2.5%~5.0% of full training dataset size, the positioning error is *E* = (3.00~5.76) × *E0,A*;

- 3 sources when fine-tuning dataset size is *x*% = 10.0%~25.0% of full training dataset size, the positioning error is *E* = (3.35~5.96) × *E0,A*;

- 3 sources when fine-tuning dataset size is *x*% = 50.0%~100.0% of full training dataset size, the positioning error is *E* = (4.50~7.71) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for drop A.

For **direct** AI/ML positioning and **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 8 sources when fine-tuning dataset size is *x*% = 1.3%~2.5% of full training dataset size, the positioning error is *E* = ( 1.8~10.18) × *E0,B*;

- 11 sources when fine-tuning dataset size is *x*% = 4.0%~8.0% of full training dataset size, the positioning error is *E* = (1.77~7.05) × *E0,B*;

- 9 sources when fine-tuning dataset size is *x*% = 10.0%~17.0% of full training dataset size, the positioning error is *E* = (1.50~5.34) × *E0,B*; 1 source the positioning error is *E* = (14.65) × *E0,B*;

- 5 sources when fine-tuning dataset size is *x*% = 20.0%~34.0% of full training dataset size, the positioning error is *E* = (1.01~1.75) × *E0,B*; 1 source the positioning error is *E =* (12.23) × *E0,B*;

- 5 sources when fine-tuning dataset size is *x*% = 50.0% of full training dataset size, the positioning error is *E* = (1.09~1.25) × *E0,B*;

- 4 sources when fine-tuning dataset size is *x*% = 95%~100.0% of full training dataset size, the positioning error is *E* = (0.82~1.84) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for clutter parameter B.

For **direct** AI/ML positioning and **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter A and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 6 sources when fine-tuning dataset size is *x*% = 2.5% of full training dataset size, the positioning error is *E* = (2.24~22.11) × *E0,A*;

- 7 sources when fine-tuning dataset size is *x*% = (5.0%~5.6%) of full training dataset size, the positioning error is *E* = (2.02~19.49) × *E0,A*;

- 6 sources when fine-tuning dataset size is *x*% = (10.0%~25.0%) of full training dataset size, the positioning error is *E* = (1.40~18.65) × *E0,A*;

- 5 sources when fine-tuning dataset size is *x*% = 50.0% of full training dataset size, the positioning error is *E* = (1.20~10.72) × *E0,A*;

- 3 sources when fine-tuning dataset size is *x*% = 95.0%~100.0% of full training dataset size, the positioning error is *E* = (2.08~12.58) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for clutter parameter A.

For **direct** AI/ML positioning and **different network synchronization error**, evaluation has been performed where the AI/ML model is (a) previously trained for network synchronization error = A (ns) with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for network synchronization error = B (ns) with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under network synchronization error = B (ns) and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 5 sources when fine-tuning dataset size is *x*% = (1.3%~2.5%) of full training dataset size, the positioning error is *E* = (0.98~5.21) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = (4.0%~8.0%) of full training dataset size, the positioning error is *E* = (0.84~10.70) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = (10.0%~25.0%) of full training dataset size, the positioning error is *E* = (0.80~10.38) × *E0,B*;

- 1 source when fine-tuning dataset size is *x*% = (50.0%~100.0%) of full training dataset size, the positioning error is *E* = (0.81~1.1) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for network synchronization error = B (ns).

For **direct** AI/ML positioning and **different network synchronization error**, evaluation has been performed where the AI/ML model is (a) previously trained for network synchronization error = 0 ns with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for network synchronization error = 50 ns with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under network synchronization error = 0 ns and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 2 sources when fine-tuning dataset size is *x*% = (2.5%~10.0%) of full training dataset size, the positioning error is *E* = (5.08~23.44) × *E0,A*;

- 1 source when fine-tuning dataset size is *x*% = (25.0%~100.0%) of full training dataset size, the positioning error is *E* = (2.28~3.92) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for network synchronization error = 0 ns.

For **direct** AI/ML positioning and **different UE timing error**, evaluation has been performed where the AI/ML model is (a) previously trained **without UE timing error** with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning **with UE timing error** with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested **with UE timing error** and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 2 sources when fine-tuning dataset size is *x*% = 1.3%~20.0% of full training dataset size, the positioning error is *E* = (0.51~2.53) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for the case **with UE timing error**.

For **direct** AI/ML positioning and **different InF scenarios**, evaluation has been performed where the AI/ML model is (a) previously trained for InF scenario A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for InF scenario B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under InF scenario B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 5 sources when fine-tuning dataset size is *x*% = (2.0%~5.6%) of full training dataset size, the positioning error is *E* = (0.5~16.67) × *E0,B*;

- 5 sources when fine-tuning dataset size is *x*% = (8.0%~15.0%) of full training dataset size, the positioning error is *E* = (0.4~12.6) × *E0,B*;

- 2 sources when fine-tuning dataset size is *x*% = 25.0% of full training dataset size, the positioning error is *E* = (1.60~1.67) × *E0,B*;

- 2 sources when fine-tuning dataset size is *x*% = (50.0%~100.0%) of full training dataset size, the positioning error is *E* = (0.92~1.41) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for InF scenario B.

For **direct** AI/ML positioning and **different InF scenarios**, evaluation has been performed where the AI/ML model is (a) previously trained for InF scenario A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for InF scenario B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under InF scenario A and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 3 sources when fine-tuning dataset size is *x*% = (2.5%~10.0%) of full training dataset size, the positioning error is *E* = (2.28~30.2) × *E0,A*;

- 2 sources when fine-tuning dataset size is *x*% = (25.0%~100.0%) of full training dataset size, the positioning error is *E* = (1.7~9.24) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for InF scenario A.

For **direct** AI/ML positioning and **different SNR value (dB)**, evaluation has been performed where the AI/ML model is (a) previously trained for SNR value A (dB) with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for SNR value B (dB) with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under SNR value B (dB) and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = (5.6%~11.1%) of full training dataset size, the positioning error is *E* = (1.60~1.90) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for SNR value B (dB).

For **direct** AI/ML positioning and **different time varying assumptions**, evaluation has been performed where the AI/ML model is (a) previously trained for the scenario without time varying change with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for the scenario with time varying change with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under the scenario with time varying change and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = (3.7%~22.0%) of full training dataset size, the positioning error is *E* = (1.68~3.49) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for the scenario with time varying change.

For **direct** AI/ML positioning and **different channel estimation error**, evaluation has been performed where the AI/ML model is (a) previously trained for channel estimation error = 20 dB with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for channel estimation error = 0 dB with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under channel estimation error = 0 dB and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = (2.5%~25.0%) of full training dataset size, the positioning error is *E* = (1.50~2.79) × *E0,B*;

- 1 source when fine-tuning dataset size is *x*% = (50.0%~100.0%) of full training dataset size, the positioning error is *E* = (0.96~1.17) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for channel estimation error = 0 dB.

For **direct** AI/ML positioning and **different channel estimation error**, evaluation has been performed where the AI/ML model is (a) previously trained for channel estimation error = 20 dB with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for channel estimation error = 0 dB with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under channel estimation error = 20 dB and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = (2.5%~25.0%) of full training dataset size, the positioning error is *E* = (4.22~5.95) × *E0,A*;

- 1 source when fine-tuning dataset size is *x*% = (50.0%~100.0%) of full training dataset size, the positioning error is *E* = (3.08~3.94) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for channel estimation error = 20 dB.

***AI/ML assisted positioning***

For AI/ML **assisted** positioning with timing information as model output and for **different drops**, evaluation has been performed where the AI/ML model is (a) previously trained for drop A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for drop B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under drop B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 2 sources when fine-tuning dataset size is *x*% = (2.0%~10.0%) of full training dataset size, the positioning error is *E =* (1.27~7.68) × *E0,B*;

- 2 sources when fine-tuning dataset size is *x*% = (12.0%~34.0%) of full training dataset size, the positioning error is *E =* (5.59~12.88) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for drop B.

For AI/ML **assisted** positioning with timing information as model output and for **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 5 sources when fine-tuning dataset size is *x*% = (2.0%~2.5%) of full training dataset size, the positioning error is *E =* (1.47~5.88) × *E0,B*;

- 6 sources when fine-tuning dataset size is *x*% = (4.0%~5.0%) of full training dataset size, the positioning error is *E =* (1.39~4.42) × *E0,B*;

- 7 sources when fine-tuning dataset size is *x*% = (8.0%~12.0%) of full training dataset size, the positioning error is *E =* (1.34~3.93) × *E0,B*;

- 3 sources when fine-tuning dataset size is *x*% = 25.0% of full training dataset size, the positioning error is *E =* (1.33~1.91) × *E0,B*;

- 3 sources when fine-tuning dataset size is *x*% = 50.0% of full training dataset size, the positioning error is *E =* (1.15~1.33) × *E0,B*;

- 2 sources when fine-tuning dataset size is *x*% = 100.0% of full training dataset size, the positioning error is *E =* (0.89~1.15) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for clutter parameter B.

For AI/ML **assisted** positioning with timing information as model output and for **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter A and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 4 sources when fine-tuning dataset size is *x*% = (2.5%~5.0%) of full training dataset size, the positioning error is *E =* (1.47~12.94) × *E0,A*;

- 5 sources when fine-tuning dataset size is *x*% = 10.0% of full training dataset size, the positioning error is *E =* (1.32~11.52) × *E0,A*;

- 3 sources when fine-tuning dataset size is *x*% = 25.0% of full training dataset size, the positioning error is *E =* (1.22~7.65) × *E0,A*;

- 3 sources when fine-tuning dataset size is *x*% = 50.0% of full training dataset size, the positioning error is *E =* (1.2~5.86) × *E0,A*;

- 2 sources when fine-tuning dataset size is *x*% = 100.0% of full training dataset size, the positioning error is *E =* (2.64~4.66) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for the clutter parameter A.

For AI/ML **assisted** positioning and **different network synchronization error**, evaluation has been performed where the AI/ML model is (a) previously trained for network synchronization error A (ns) with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for network synchronization error B (ns) with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under network synchronization error B (ns) and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 5 sources when fine-tuning dataset size is *x*% = 2.0%~5.0% of full training dataset size, the positioning error is *E =* (1.28~5.44) × *E0,B*;

- 5 sources when fine-tuning dataset size is *x*% = 8.0%~25.0% of full training dataset size, the positioning error is *E =* (1.10~4.07) × *E0,B*;

- 1 source when fine-tuning dataset size is *x*% = 50.0%~100.0% of full training dataset size, the positioning error is *E =* (1.01~1.47) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for network synchronization error B (ns).

For AI/ML **assisted** positioning and **different network synchronization error**,

- evaluation has been performed where the AI/ML model is (a) previously trained for network synchronization error = 0 ns with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for network synchronization error = 50 ns with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under network synchronization error = 0 ns and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that, denoting *E0,A* (meters) as the full training accuracy at CDF=90% for network synchronization error = 0 ns,

- 2 sources when fine-tuning dataset size is *x*% = (2.5%~100.0%) of full training dataset size, the positioning error is *E =* (3.71~5.97) × *E0,A*;

- evaluation has been performed where the AI/ML model is (a) previously trained for network synchronization error = 50 ns with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for network synchronization error = 0 ns with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under network synchronization error = 50 ns and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that, denoting *E0,A* (meters) as the full training accuracy at CDF=90% for network synchronization error = 50 ns,

- 1 source when fine-tuning dataset size is *x*% = (2.5%~100.0%) of full training dataset size, the positioning error is *E =* (1.15~2.23) × *E0,A*;

For AI/ML **assisted** positioning and **different InF scenarios**, evaluation has been performed where the AI/ML model is (a) previously trained for InF scenario A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for InF scenario B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under InF scenario B and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 3 sources when fine-tuning dataset size is *x*% = (2.0%~12.0%) of full training dataset size, the positioning error is *E =* (1.20~6.0) × *E0,B*;

- 1 source when fine-tuning dataset size is *x*% = 25.0%~50.0% of full training dataset size, the positioning error is *E =* (2.55~2.91) × *E0,B*;

Here *E0,B* (meters) is the full training accuracy at CDF=90% for InF scenario B.

For AI/ML **assisted** positioning and **different InF scenarios**, evaluation has been performed where the AI/ML model is (a) previously trained for InF-DH{60%,6m,2m} with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for InF-SH{20%,2m,10m} with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under InF-DH{60%,6m,2m} and the horizontal accuracy at CDF=90% is *E* meters. Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = 2.5%-50.0% of full training dataset size, the positioning error is *E =* (2.53~3.44) × *E0,A*;

Here *E0,A* (meters) is the full training accuracy at CDF=90% for InF-DH{60%,6m,2m}.

For AI/ML **assisted** positioning with LOS/NLOS indicator as model output and for **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter B and the LOS/NLOS indication accuracy is *E* (using F1-score). Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = 10.0% of full training dataset size, the accuracy (using F1-score) of LOS/NLOS indicator is *E =* (0.56~0.974) × *E0,B*;

Here *E0,B* is the full training accuracy (using F1-score) for the clutter parameter B.

For AI/ML **assisted** positioning with LOS/NLOS indicator as model output and for **different clutter parameters**, evaluation has been performed where the AI/ML model is (a) previously trained for clutter parameter A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for clutter parameter B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under clutter parameter A and the LOS/NLOS indication accuracy is *E* (using F1-score). Evaluation results show that,

- 1 source when fine-tuning dataset size is *x*% = 10.0% of full training dataset size, the accuracy (using F1-score) of LOS/NLOS indicator is *E =* (0.09~0.24) × *E0,A*;

Here *E0,A* is the full training accuracy (using F1-score) for the clutter parameter A.

***Both direct AI/ML positioning and AI/ML assisted positioning***

As a summary of the observations above, for both direct AI/ML positioning and AI/ML assisted positioning, evaluation results show that:

- Fine-tuning/re-training a previous model with dataset of the new deployment scenario improves the model performance for the new deployment scenario. For details on the amount of improvement, see the observations listed above.

- After fine-tuning/re-training a previous model with dataset of the new deployment scenario, the performance of the updated model degrades for the previous deployment scenario (e.g., previous clutter parameter setting) that the previous model was trained for.

- Examples of the deployment scenario include: different drops, different clutter parameter, different InF scenarios

For both direct AI/ML positioning and AI/ML assisted positioning,

- if the new deployment scenario is significantly different from the previous deployment scenario the model was trained for (e.g., different drops, different clutter parameter, different InF scenarios), fine-tuning a previous model requires similarly large training dataset size as training the model from scratch, in order to achieve the similar performance for the new deployment scenario.

- If the new deployment scenario is NOT significantly different from the previous deployment scenario the model was trained for (e.g., 2ns difference in network synchronization error between the previous and the new deployment scenario), fine-tuning a previous model requires a small (e.g., *x*%=10%) training dataset size as compared to training the model from scratch, in order to achieve the similar performance for the new deployment scenario.

#### 6.4.2.4 Model-input Size Reduction

***Observations*:**

***Direct AI/ML positioning***

For the evaluation of direct AI/ML positioning, with Nt consecutive time domain samples used as model input, evaluation results show that when CIR, PDP, or DP is used as model input, using different Nt while holding other parameters the same,

- Reducing Nt from 256 to 128 does not appreciably degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/2 that of Nt=256.

- Positioning error of Nt=128 is 0.81 ~ 1.19 times the positioning error of Nt=256;

- Reducing Nt from 256 to 64~32 may degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/4 ~1/8 that of Nt=256, respectively.

- Positioning error of Nt=64 is 0.88 ~ 3.00 times the positioning error of Nt=256;

- Positioning error of Nt=32 is 1.05 ~ 4.29 times the positioning error of Nt=256;

- Note: the variation in the positioning accuracy depends on each company's simulation assumption (e.g., AI/ML complexity).

For direct AI/ML positioning, the evaluation of positioning accuracy at model inference is affected by the type of model input and AI/ML complexity. For a given AI/ML model design, there is a tradeoff between model input, AI/ML complexity (model complexity and computational complexity), and positioning accuracy. Evaluation results show that if changing model input type while holding other parameters (e.g., Nt, N't, Nport, N'TRP) the same,

- When comparing PDP and CIR as model input,

- 9 sources showed evaluation results where the positioning error of PDP as model input is 1.06 ~ 1.62 times the positioning error of CIR as model input.

- 5 sources showed evaluation results where the positioning error of PDP as model input is 0.61 ~ 0.96 times the positioning error of CIR as model input.

- When comparing DP and CIR as model input,

- 4 sources showed evaluation results where the positioning error of DP as model input is 1.18 ~ 1.96 times the positioning error of CIR as model input.

- 2 sources showed evaluation results where the positioning error of DP as model input is 0.79~0.92 times the positioning error of CIR as model input.

- Note: For one of the sources (R1-2306112), the difference in relative performance is due to the complexity of the AI/ML model.

- Note: For another source (R1-2307920), the difference in relative performance is due to the parameter settings.

- Note: the variation in the positioning accuracy depends on each company's simulation assumption (e.g., AI/ML complexity).

For the evaluation of direct AI/ML positioning, when N't time domain samples with the strongest power are selected as model input, evaluation results show that:

- For model input of CIR or PDP and Nt=256, using different N't while holding other parameters constant,

- Reducing N't from 256 to 64 does not appreciably degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/4 that of Nt=N't=256.

- Positioning error of N't=128 is 1.02 ~ 1.07 times the positioning error of Nt=N't=256;

- Positioning error of N't=64 is 1.02 ~ 1.21 times the positioning error of Nt=N't=256;

- Reducing N't from 256 to 32~16 degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/8 ~ 1/16 that of Nt=N't=256.

- Positioning error of N't=32 is 1.14 ~ 2.03 times the positioning error of Nt=N't=256;

- Positioning error of N't=16 is 1.12 ~ 2.54 times the positioning error of Nt=N't=256;

- Reducing N't from 256 to 9~8 degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/32 that of Nt=N't=256.

- Positioning error of N't=9~8 is 1.42 ~ 3.29 times the positioning error of Nt=N't=256;

- For model input of DP and Nt=256, using different N't while holding other parameters constant,

- One source (R1-2304339) showed that reducing N't from 64 to 32 does not degrade the positioning accuracy while the measurement size and signalling overhead shrink by (approximately) 1/2.

- Positioning error of N't=32 is 1.03 times the positioning error of N't=64.

- Note: the evaluation results based on the other model input (e.g., multiple path) can be added in next meeting

Based on evaluation results by 8 sources, for TRP reduction of **direct** AI/ML positioning, approaches supporting dynamic TRP pattern can achieve the horizontal positioning accuracy *Edynamic* = (0.80~2.15) × *Efixed* (meters), when other design parameters are held the same, where:

*- Edynamic* (meters) is the horizontal positioning accuracy at CDF=90% for approaches supporting dynamic TRP pattern (i.e., Approach 1-B and 2-B);

*- Efixed* (meters) is the horizontal positioning accuracy at CDF=90% for approaches supporting fixed TRP pattern (i.e., Approach 1-A and 2-A);

Based on evaluation results by 8 sources, for TRP reduction of **direct** AI/ML positioning, Approach 1-A and 2-A achieve similar performance. The horizontal positioning accuracy *E*2A = (0.87~1.32) × *E*1A (meters), when other design parameters are held the same, where:

*- E*1A (meters) is the horizontal positioning accuracy at CDF=90% for Approach 1-A;

*- E*2A (meters) is the horizontal positioning accuracy at CDF=90% for Approach 2-A;

Based on evaluation results by 11 sources, for TRP reduction of **direct** AI/ML positioning, the positioning accuracy degrades as the number of active TRPs are reduced from 18 TRPs to 3 TRPs. The degradation increases as the number of active TRPs decreases.

- When the number of active TRP is reduced from NTP = 18 to N'TP = 12~8, the average horizontal positioning accuracy *E* is in the range of *E* = (1.48~1.95) × *E*18TRP;

- When the number of active TRP is reduced from NTP = 18 to N'TP = 6~5, the average horizontal positioning accuracy *E* is in the range of *E* = (2.35~3.04) × *E*18TRP;

- When the number of active TRP is reduced from NTP = 18 to N'TP = 4~3, the average horizontal positioning accuracy *E* is in the range of *E* = (2.13~5.11) × *E*18TRP;

Here *E* (meters) is the horizontal positioning accuracy at CDF=90% with N'TP active TRPs, *E*18TRP (meters) is the horizontal positioning accuracy at CDF=90% with NTP = 18 active TRPs.

Note: some results from 2 sources show *E* > 11 × *E*18TRP for N'TP= 9 and 6 when using Approach 2-B.

***AI/ML assisted positioning***

For AI/ML assisted positioning, the positioning accuracy at model inference is affected by the type of model input. Evaluation results show that if changing model input type while holding other parameters (e.g., Nt, N't, Nport, N'TRP) the same,

- The positioning error of PDP as model input is 1.17 ~ 1.63 times the positioning error of CIR as model input.

- The positioning error of DP as model input is 1.33 ~ 2.01 times the positioning error of CIR as model input.

For AI/ML assisted positioning, with Nt consecutive time domain samples used as model input, evaluation results show that when CIR or PDP are used as model input, using different Nt while holding other parameters the same,

- Reducing Nt from 256 to 128 does not appreciably degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/2 that of Nt=256.

- Positioning error of Nt=128 is 1.00 ~ 1.42 times the positioning error of Nt=256;

- Reducing Nt from 256 to 64~32 may degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/4 ~1/8 that of Nt=256, respectively.

- Positioning error of Nt=64 is 1.09 ~ 3.02 times the positioning error of Nt=256;

- Positioning error of Nt=32 is 2.43 ~ 5.10 times the positioning error of Nt=256;

For AI/ML assisted positioning, when N't time domain samples with the strongest power are selected as model input, evaluation results show that for model input of CIR or PDP and Nt=256, using different N't while holding other parameters the same,

- Reducing N't from 256 to 64 does not appreciably degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/4 that of Nt=N't=256.

- Positioning error of N't=128 is 1.00 ~ 1.33 times the positioning error of Nt=N't=256;

- Positioning error of N't=64 is 0.98 ~ 1.23 times the positioning error of Nt=N't=256;

- Reducing N't from 256 to 32~16 may degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/8 ~ 1/16 that of Nt=N't=256.

- Positioning error of N't=32 is 1.15 ~ 1.69 times the positioning error of Nt=N't=256;

- Positioning error of N't=16 is 1.04 ~ 2.67 times the positioning error of Nt=N't=256;

- Reducing N't from 256 to 9 degrade the positioning accuracy, while the measurement size and signalling overhead shrink to (approximately) 1/32 that of Nt=N't=256.

- Positioning error of N't=9 is 1.66 ~ 4.40 times the positioning error of Nt=N't=256;

Based on evaluation results by 2 sources, for TRP reduction of AI/ML **assisted** positioning with multi-TRP construction, approaches supporting dynamic TRP pattern can achieve the horizontal positioning accuracy *Edynamic* = (1.03~1.74) × *Efixed* (meters), when other design parameters are held the same, where:

*- Edynamic* (meters) is the horizontal positioning accuracy at CDF=90% for approaches supporting dynamic TRP pattern (i.e., Approach 1-B and 2-B);

*- Efixed* (meters) is the horizontal positioning accuracy at CDF=90% for approaches supporting fixed TRP pattern (i.e., Approach 1-A and 2-A);

Note: evaluation results of 1 source show *Edynamic* = (5.66~8.12) × *Efixed* when the number of active TRP is reduced from NTP =18 to N'TP =9 or 4.

Based on evaluation results by 2 sources, for TRP reduction of AI/ML **assisted** positioning, Approach 1-A and 2-A achieve similar performance. The horizontal positioning accuracy *E*2A = (1~1.47) × *E*1A (meters), when other design parameters are held the same, where:

*- E*1A (meters) is the horizontal positioning accuracy at CDF=90% for Approach 1-A;

*- E*2A (meters) is the horizontal positioning accuracy at CDF=90% for Approach 2-A;

Based on evaluation results by 4 sources, for TRP reduction of AI/ML **assisted** positioning, the positioning accuracy degrades as the number of active TRPs are reduced from 18 TRPs to 3 TRPs. The degradation increases as the number of active TRPs decreases.

- When the number of active TRP is reduced from NTP =18 to N'TP =9, the average horizontal positioning accuracy is *E* = 2.01 × *E*18TRP;

- When the number of active TRP is reduced from NTP =18 to N'TP = 6, the average horizontal positioning accuracy is *E* = 3.04 × *E*18TRP;

- When the number of active TRP is reduced from NTP =18 to N'TP = 3~4, the average horizontal positioning accuracy is *E* = (5.01~6.53) × *E*18TRP;

Here *E* (meters) is the horizontal positioning accuracy at CDF=90% with N'TP active TRPs, *E*18TRP (meters) is the horizontal positioning accuracy at CDF=90% with NTP =18 active TRPs.

Note: some results from 1 source show *E* > 7.54 × *E*18TRP for N'TP=9 and *E* > 42.76 × *E*18TRP for N'TP=6 when using Approach 1-B/2-B.

***Both direct AI/ML positioning and AI/ML assisted positioning***

Evaluation of TRP reduction for **both** direct AI/ML positioning and AI/ML assisted positioning shows that: identification of the active TRPs is beneficial for Approach 2-B. Otherwise, the model suffers from poor performance in terms of positioning accuracy.

For example, evaluation results from 4 sources show that the horizontal positioning accuracy is greater than 10 m if TRP identification is not included as model input.

#### 6.4.2.5 Non-ideal label(s)

***Observations*:**

***Direct AI/ML positioning***

For direct AI/ML positioning, for L in the range of 0.25m to 5m, the positioning error increases approximately in proportion to L, where L (in meters) is the standard deviation of truncated Gaussian Distribution of the ground-truth label error.

Evaluation shows that direct AI/ML positioning is robust to certain *label error* based on evaluation results of L in the range of (0, 5) meter. The exact range of label error that can be tolerated depends on the positioning accuracy requirement, where tighter positioning accuracy requirement demands smaller label error.

***AI/ML assisted positioning***

In evaluation of AI/ML assisted positioning with timing information (e.g., TOA) as model output, for L in the range of 0.25m to 5m, the timing (e.g., TOA) estimation error and positioning error increases approximately in proportion to L, where L (in meters) is the standard deviation of truncated Gaussian distribution of the ground-truth label error.

Evaluations show that AI/ML assisted positioning with timing information (e.g., ToA) as *model output* is robust to certain *label error* based on evaluation results of L in the range of (0, 5) meter. The exact range of label error that can be tolerated depends on the positioning accuracy requirement, where tighter positioning accuracy requirement demands smaller label error.

Based on evaluation results from 3 sources, for AI/ML assisted positioning where the model output includes the LOS/NLOS indicator, when the model is trained with dataset containing random LOS/NLOS label error, the models have no or minor degradation for LOS/NLOS identification accuracy up to at least m%=20% and at least n%=20%. When the training dataset has up to m%=20% and n%=20%, evaluation results show that the **LOS/NLOS identification accuracy** is PlablErr = PnoLablErr – d (percentage), where d is in the range of (−1.2%~3.1%).

- PnoLablErr (percentage) is the LOS/NLOS identification accuracy when m%=0% and n%=0%;

***Other***

For AI/ML based positioning, evaluation results show that semi-supervised learning is helpful for improving the positioning accuracy when the same amount of ideal labelled data is used for supervised learning, and the number of ideal labelled data is limited.

Regarding ground-truth label generation for AI/ML based positioning, multiple sources submitted evaluation results on the impact of ground-truth label for training obtained by existing NR RAT-dependent positioning methods. Feasibility and performance benefit of utilizing ground-truth label for training estimated by existing NR RAT-dependent positioning methods are observed.

- Source 1 evaluated in InF-DH {40%, 2, 2} and showed that AI/ML model can be trained with noisy labels along with the corresponding quality estimated by the legacy positioning methods, to improve positioning performance from 3.73m@90% (5k ideal label) to 1.72m @90% (5k ideal label + 20k noisy label). It also showed that the performance benefit compared to semi-supervised training of 2.78m @90% (5k ideal label + 20k unlabeled data). Note that training data weighting is used with label quality indicator.

- Source 2 evaluated in InF-DH {60%, 6, 2} and showed that the performance of direct AI/ML positioning with 1k clean labelled samples improves from 13.76m to 8.72m when considering additional 350 samples that are labelled using NR-RAT positioning method. Note that the label error is up to 3.5m.

- Source 3 evaluated in both InF-DH {60%, 6, 2} and InF-DH {40%, 2, 2} and showed performance loss when compared to all ideal label case. For example it showed in InF-DH {40%, 2, 2} the accuracy degrades from 0.39m @90% (100% ideal label) to 2.10m @90% (50% ideal label and 50% label obtained by existing DL-TDOA scheme). Note that noisy label is treated the same as ideal label in training.

#### 6.4.2.6 Summary of Performance Results for Positioning accuracy enhancements

For the use case of positioning accuracy enhancement, extensive evaluations have been carried out. Both direct AI/ML positioning and AI/ML assited positioning are evaluated using one-sided model. The following areas are investigated.

- **Performance evaluation without generalization consideration**, where the AI/ML model is trained and tested with dataset of the same deployment scenario.

o **AI/ML vs RAT-dependent positioning methods**. For the basic performance without generalization consideration, AI/ML based positioning can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods. For example, in InF-DH with clutter parameter setting {60%, 6m, 2m}, AI/ML based positioning can achieve horizontal positioning accuracy of <1m at CDF=90%, as compared to >15m for conventional positioning method.

o **Impact of training data sample density** (i.e., training dataset size for a given evaluation area). Evaluation with uniform UE distribution shows that, the larger the training dataset size (i.e., higher sample density), the smaller the positioning error (in meters), until a saturation point is reached where additional training data does not bring further improvement to the positioning accuracy.

- **AI/ML complexity**. For a given company’s model design, in terms of model inference complexity (model complexity and computational complexity), a lower complexity model can still achieve acceptable positioning accuracy (e.g., <1m), albeit degraded, when compared to a higher complexity model.

- **Generalization study**. Evaluations are carried out to investigate various generalization aspects, where the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario. The generalization aspects include: different drops; different clutter parameters; different InF scenarios; network synchronization error; UE/gNB RX and TX timing error; SNR mismatch; channel estimation error; time varying changes.

Methods are evaluated which have been shown to be able to handle generalization issues, including:

o **Better training dataset construction (i.e., mixed dataset)**, where the training dataset is composed of data from multiple deployment scenarios, which include data from the same deployment scenario as the test dataset.

o **Fine-tuning/re-training**, where the model is re-trained/fine-tuned with a dataset from the same deployment scenario as the test dataset. The impact of the amount of fine-tuning data on the positioning accuracy of the fine-tuned model is evaluated for the various generalization aspects. Evaluation results are obtained for two experiments:

 The AI/ML model is (a) previously trained for scenario A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for scenario B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under scenario B. The horizontal positioning accuracy at CDF=90% is *E* meters.

 The AI/ML model is (a) previously trained for scenario A with a dataset of sample density *N* (#samples/m2), (b) followed by fine-tuning for scenario B with a dataset of sample density *x*% × *N* (#samples/m2), (c) then tested under scenario A. The horizontal positioning accuracy at CDF=90% is *E* meters.

- **Model input size reduction**. Evaluations are carried out to examine various ways to change the model input size and its impact on positioning accuracy:

o Different measurement type, for example, CIR, PDP, DP.

o Different number of consecutive time domain samples, Nt.

o Different number of non-zero samples N't selected from the Nt consecutive time domain samples (N't < Nt).

o Different number of active TRPs, N'TRP.

The model input size for various measurement type (CIR, PDP, DP) and dimensions (N'TRP, Nt, N't, Nport) is analyzed. Evaluation results show that, model input of different measurement type and dimensions can have different reporting overhead and positioning accuracy.

- **Fixed TRP pattern vs dynamic TRP pattern**. Evaluation results show that, approaches supporting dynamic TRP pattern may be able to achieve comparable horizontal positioning accuracy as approaches supporting fixed TRP pattern, when other design parameters are held the same.

- **Model output of AI/ML assisted positioning**. For AI/ML assisted positioning, evaluations are carried out where the model output includes timing information and/or LOS/NLOS indicator, in the format of hard- or soft- value.

- **Non-ideal label in the training dataset**. Evaluations are carried out to show the impact of:

o Label error, where the label in the training dataset is degraded from ground-truth label by an error.

 For direct AI/ML positioning and AI/ML assisted positioning with timing information as model output, location error in each dimension of x-axis and y-axis is modelled as a truncated Gaussian distribution.

 For AI/ML assisted positioning where the model output includes the LOS/NLOS indicator, random LOS/NLOS label error is applied.

o Absent label, where some data samples in the training dataset do not have associated labels. Semi-supervised learning is evaluated for this case.

- **Model monitoring**. Preliminary evaluation of model monitoring methods are provided by individual companies. The following methods are shown to be feasible:

o Label based methods, where ground-truth label (or its approximation) is provided for monitoring the accuracy of model output.

o Label-free methods, where model monitoring does not require ground-truth label (or its approximation).

Based on RAN1 evaluations of AI/ML based positioning,

- It is beneficial to support both direct AI/ML and AI/ML assisted positioning approaches since they can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods in the evaluated indoor factory scenarios.

- Both UE-side model and NW-side model can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods.

- It is desired to apply methods to handle generalization aspects.

- It is desired to consider training data collection requirements.

- If AI/ML based positioning is considered for normative work, it is desired to further investigate model input design aspects: the model input type (e.g., CIR, PDP, DP), dimension (e.g., parameters N'TRP, Nt, N't, Nport) and related format (e.g., for the timing information: absolute time or relative time) considering the trade-off of positioning accuracy, signalling overhead, and AI/ML complexity.

# 7 Potential specification impact assessment

## 7.1 Physical layer aspects

In this clause, aspects related to, e.g., the potential specification of the AI Model lifecycle management, and dataset construction for training, validation and test for the selected use cases are considered.

In addition, use case and collaboration level specific specification impact is documented, such as new signalling, means for training and validation data assistance, assistance information, measurement, and feedback.

### 7.1.1 Common framework

***Items considered for studying the necessity, feasibility, potential specification impact:***

*Performance monitoring*

The following metrics/methods for AI/ML model monitoring in lifecycle management per use case are considered:

- Monitoring based on inference accuracy, including metrics related to intermediate KPIs

- Monitoring based on system performance, including metrics related to system peformance KPIs

- Other monitoring solutions, at least the following 2 options.

- Monitoring based on data distribution

- Input-based: e.g., Monitoring the validity of the AI/ML input, e.g., out-of-distribution detection, drift detection of input data, or SNR, delay spread, etc.

- Output-based: e.g., drift detection of output data

- Monitoring based on applicable condition

Note: Monitoring metric calculation may be done at NW or UE

Methods to assess/monitor the applicability and expected performance of an *inactive model/functionality*, including the following examples for the purpose of activation/selection/switching of UE-side models/UE-part of two-sided models /functionalities (if applicable):

- Assessment/Monitoring based on the additional conditions associated with the model/functionality

- Assessment/Monitoring based on input/output data distribution

- Assessment/Monitoring using the inactive model/functionality for monitoring purpose and measuring the inference accuracy

- Assessment/Monitoring based on past knowledge of the performance of the same model/functionality (e.g., based on other UEs)

### 7.1.2 CSI feedback enhancement

***Items considered for studying the necessity, feasibility, potential specification impact:***

**In CSI compression using two-sided model use case:**

*Performance monitoring:*

- Model performance monitoring related assistance signalling and procedure.

- Metrics/methods including:

- Intermediate KPIs (e.g., SGCS)

- Eventual KPIs (e.g., Throughput, hypothetical BLER, BLER, NACK/ACK).

- Legacy CSI based monitoring: schemes using additional legacy CSI reporting

- Other monitoring solutions, at least including the following option:

- Input or Output data based monitoring: such as data drift between training dataset and observed dataset and out-of-distribution detection

- NW-side performance monitoring: NW monitors the performance and make decisions of model/functionality activation/ deactivation/updating/switching. Impact to enable performance monitoring using an existing CSI feedback scheme as the reference, including the association between AI/ML scheme and existing CSI feedback scheme for monitoring, are considered. Note: The metric for monitoring and comparison includes intermediate KPI and eventual KPI.

- UE-side performance monitoring: UE monitors the performance and reports to Network, NW makes decisions of model/functionality activation/deactivation/updating/switching. Impact on triggering and means for reporting the monitoring metrics, including periodic/semi-persistent and aperiodic reporting, and other reporting initiated from UE, are not precluded.

*Intermediate KPI based model monitoring:*

The following intermediate KPI-based model monitoring options were proposed by companies:

- NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE or obtained from the UE-side.

- UE-side monitoring based on the output of the CSI reconstruction model, subject to the aligned format, associated to the CSI report, indicated by the NW or obtained from the network side.

- Network may configure a threshold criterion to facilitate UE to perform model monitoring.

- UE-side monitoring based on the output of the CSI reconstruction model at the UE-side

- Note: CSI reconstruction model at the UE-side can be the same or different comparing to the actual CSI reconstruction model used at the NW-side. Network may configure a threshold criterion to facilitate UE to perform model monitoring.

*Fallback mode:*

- Potential specification impact for supporting co-existence and fallback mechanisms between AI/ML-based CSI feedback mode and legacy non-AI/ML-based CSI feedback mode

*NW/UE alignment:*

- Alignment of the quantization/dequantization method and the feedback message size between Network and UE, including the following:

- For vector quantization scheme, the format and size of the VQ codebook, and the size and segmentation method of the CSI generation model output

- For scalar quantization scheme, uniform and non-uniform quantization with format, e.g., quantization granularity, consisting of distribution of bits assigned to each float.

- Quantization alignment for CSI feedback between CSI generation part at the UE and CSI reconstruction part at the NW is needed, e.g.,

 - through model pairing process,

 - alignment based on standardized quantization scheme.

 - Additional methods not precluded.

*Model input/output:*

- Output-CSI-UE and input-CSI-NW at least for Precoding matrix

- Option 1a: The precoding matrix in spatial-frequency domain

- Option 1b: The precoding matrix represented using angular-delay domain projection

- whether Option 2: Explicit channel matrix (i.e., full Tx \* Rx MIMO channel) is also studied depends on the performance evaluations:

- Option 2a: raw channel is in spatial-frequency domain

- Option 2b: raw channel is in angular-delay domain

- CSI part 1 includes at least CQI for first codeword, RI, and information representing the part 2 size. CSI part 2 includes at least the content of CSI generation part output. Other CSI report formats are not precluded.

*UE side data collection:*

- Enhancement of CSI-RS configuration to enable higher accuracy measurement.

- Assistance information for UE data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc.

- The provision of assistance information needs to consider feasibility of disclosing proprietary information to the other side.

- Signalling for triggering the data collection

*NW side data collection:*

- Enhancement of SRS and/or CSI-RS measurement and/or CSI reporting to enable higher accuracy measurement.

- Contents of the ground-truth CSI including:

- Data sample type, e.g., precoding matrix, channel matrix etc.

- Data sample format: scaler quantization and/or codebook-based quantization (e.g., e-type II like).

- Assistance information (e.g., time stamps, and/or cell ID, Assistance information for Network data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc., and data quality indicator)

- Latency requirement for data collection

- Signalling for triggering the data collection

- Ground-truth CSI report for NW side data collection *for model performance monitoring*, including:

- Scalar quantization for ground-truth CSI

- Codebook-based quantization for ground-truth CSI

- RRC signalling and/or L1 signalling procedure to enable fast identification of AI/ML model performance

 Aperiodic/semi-persistent or periodic ground-truth CSI report

- Ground-truth CSI format *for model training*, including scalar or codebook-based quantization for ground-truth CSI. The number of layers for which the ground-truth data is collected, and whether UE or NW determine the number of layers for ground-truth CSI data collection, are considered.

In CSI compression using two-sided model use case with training collaboration Type 3, for sequential training, at least the following aspects have been identified for dataset delivery from RAN1 perspective, including:

- Dataset and/or other information delivery from UE side to NW side, which can be used at least for CSI reconstruction model training

- Dataset and/or other information delivery from NW side to UE side, which can be used at least for CSI generation model training

- Potential dataset delivery methods including offline delivery, and over the air delivery

- Data sample format/type

- Quantization/de-quantization related information

*CSI configuration and report:*

- NW configuration to determine CSI payload size, e.g., possible CSI payload size, possible rank restriction and/or other related configuration.

- How UE determines/reports the actual CSI payload size and/or other CSI related information within constraints configured by the network.

- Relevant UCI format considering the legacy CSI reporting principle with CSI Part 1 and Part 2 as a starting point, where Part 1 has a network configured fixed size and Part 2 size is dynamic, determined by information in Part 1.

For CQI determination in CSI report, if CQI in CSI report is configured.

- Option 1: CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation, including

- Option 1a: CQI is calculated based on target CSI with realistic channel measurement

- Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment

- Option 1c: CQI is calculated based on legacy codebook

- Option 2: CQI is calculated based on the output of CSI reconstruction part from the realistic channel estimation, including

- Option 2a: CQI is calculated based on CSI reconstruction output, if CSI reconstruction model is available at the UE and UE can perform reconstruction model inference with potential adjustment

- Note: CSI reconstruction part at the UE can be different comparing to the actual CSI reconstruction part used at the NW.

- Option 2b: CQI is calculated using two stage approach, UE derive CQI using precoded CSI-RS transmitted with a reconstructed precoder.

- Notes: feasibility of different options should be evaluated. Gap analyses between the UE side CQI calculation results and the NW side results, as well as the impact on the scheduling performance should be evaluated. Complexity of CQI calculation needs to be evaluated, including the computing complexity and potential RS/signalling overhead.

Feasibility and methods to support the legacy CSI reporting principles:

- The priority rule regarding CSI collision handling and CSI omission

- Codebook subset restriction

- Input-CSI-NW/output-CSI-UE considered in angular-delay domain, beam restriction can be based on legacy SD basis vector-based input CSI in angular domain.

- CSI processing Unit

*Potential specification enhancement on:*

- CSI-RS configurations (not including CSI-RS pattern design enhancements)

- CSI configuration

- For network to indicate CSI reporting related information, e.g., gNB indication to the UE of one or more of following:

- Information indicating CSI payload size

- Information indicating quantization method/granularity

- Rank restriction

- Other payload related aspects

- CSI reporting configurations

- For UE determination/reporting of the actual CSI payload size, UE reports related information as configured by the NW

- CSI report UCI mapping/priority/omission

- CSI processing procedures

In CSI compression using two-sided model use case, feasibility and procedure to align the information that enables the UE to select a CSI generation model(s) compatible with the CSI reconstruction model(s) used by the gNB is studied. At least the following options have been proposed by companies to define the pairing information used to enable the UE to select a CSI generation model(s) that is compatible with the CSI reconstruction model(s) used by the gNB:

- Option 1: The pairing information is in the forms of the CSI reconstruction model ID that NW will use.

- Option 2: The pairing information is in the forms of the CSI generation model ID that the UE will use.

- Option 3: The pairing information is in the forms of the paired CSI generation model and CSI reconstruction model ID.

- Option 4: The pairing information is in the forms of by the dataset ID during type 3 sequential training.

- Option 5: The pairing information is in the forms of a training session ID to a prior training session (e.g., API) between NW and UE.

- Option 6: The pairing information is up to UE/NW offline co-engineering alignment, transparent to 3GPP specification.

- Note: the disclosure of the vendor information during the model pairing procedure and model identification procedure should be considered.

- Note: If each UE side model is compatible with all NW side model, the information is not needed for the UE.

- Note: Above does not imply there is a need for a central entity for defining/storing/maintaining the IDs.

**In CSI prediction using UE-sided model use case:**

*Data collection:*

In CSI prediction using UE sided model use case, at least the following aspects have been proposed by companies on data collection, including:

- Signalling and procedures for the data collection

- Data collection indicated by NW

- Requested from UE for data collection

- CSI-RS configuration

- Assistance information for categorizing the data, if needed

- The provision of assistance information needs to consider feasibility of disclosing proprietary information to the other side.

*Performance monitoring:*

For CSI prediction using UE side model use case, at least the following aspects have been proposed by companies on performance monitoring for functionality-based LCM:

- Type 1:

- UE calculates the performance metric(s)

- UE reports performance monitoring output that facilitates functionality fallback decision at the network

- Performance monitoring output details can be further defined

- NW may configure threshold criterion to facilitate UE side performance monitoring (if needed).

- NW makes decision(s) of functionality fallback operation (fallback mechanism to legacy CSI reporting).

- Type 2:

- UE reports *predicted CSI* and/or the corresponding *ground-truth*

- NW calculates the *performance metrics*.

- NW makes decision(s) of functionality fallback operation (fallback mechanism to legacy CSI reporting).

- Type 3:

- UE calculates the *performance metric(s)*

- UE reports *performance metric(s)* to the NW

- NW makes decision(s) of functionality fallback operation (fallback mechanism to legacy CSI reporting).

- Functionality selection/activation/deactivation/switching as defined for other UE side use cases can be reused, if applicable.

- Configuration and procedure for performance monitoring

- CSI-RS configuration for performance monitoring

- Performance metric including at least intermediate KPI (e.g., NMSE or SGCS)

- UE report, including periodic/semi-persistent/aperiodic reporting, and event driven report

- Note: down selection is not precluded.

- Note: UE may make decision within the same functionality on model selection, activation, deactivation, switching operation transparent to the NW.

### 7.1.3 Beam management

***Items considered for studying the necessity, feasibility, potential specification impact***:

*Performance monitoring:*

For the performance monitoring of BM-Case1 and BM-Case2:

- Performance metric(s) with the following alternatives:

- Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy

- Alt.2: Link quality related KPIs, e.g., throughput, L1-RSRP, L1-SINR, hypothetical BLER

- Alt.3: Performance metric based on input/output data distribution of AI/ML

- Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP

- Benchmark/reference for the performance comparison, including:

- Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)

- Alt.4: Measurements of the predicted best beam(s) corresponding to model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K Beams)

- Signalling/configuration/measurement/report for model monitoring, e.g., signalling aspects related to assistance information (if supported), Reference signals

For BM-Case1 and BM-Case2 with a UE-side AI/ML model:

- Type 1 performance monitoring:

- Configuration/Signalling from gNB to UE for measurement and/or reporting

- UE may have different operations

- Option 1 (NW-side performance monitoring): UE sends reporting to NW (e.g., for the calculation of performance metric at NW)

- Option 2 (UE-assisted performance monitoring): UE calculates performance metric(s), either reports it to NW or reports an event to NW based on the performance metric(s)

- Indication from NW for UE to do LCM operations

- Note: At least the performance and reporting overhead of model monitoring mechanism should be considered

- Type 2 performance monitoring:

- Indication/request/report from UE to gNB for performance monitoring

- Note: The indication/request/report may be not needed in some case(s)

- Configuration/Signalling from gNB to UE for performance monitoring measurement and/or reporting

- If it is for UE side model monitoring, UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation

- Mechanism that facilitates the UE to detect whether the functionality/model is suitable or no longer suitable

For BM-Case1 and BM-Case2 with a NW-side AI/ML model

- Beam measurement and report for model monitoring

- UE reporting of beam measurement(s) based on a set of beams indicated by gNB.

- Signalling, e.g., RRC-based, L1-based.

- Note: This may or may not have specification impact.

- NW monitors the performance metric(s) and makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation

- Note: Performance and UE complexity, power consumption should be considered.

Table 7.2.3-1 summarizes applicability of various alternatives for performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2.

Table 7.2.3-1: Alternatives for Performance metric(s) of AI/ML model monitoring
for BM-Case 1 and BM-Case 2

|  |  |  |  |
| --- | --- | --- | --- |
| Alt. 1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy | Alt. 2: Link quality related KPIs, .e.g., throughput, L1-RSRP, L1-SINR, hypothetical BLER | Alt.3: Performance metric based on input/output data distribution of AI/ML | Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP |
| Applicable to all studied AI models  | Applicable to all studied AI models  | Applicable to all studied AI models | May not applicable to some implementation of AI model (e.g., not output of predicted L1-RSRP) |
| Reflect the prediction accuracy of AI model | Reflect the system/link performance | Reflect the change of the statics of the input/output data  | Reflect accuracy of the predicted 1-RSRP |
| Not reflect the system/link performance directly | Not reflect the prediction accuracy of AI model directly | Not reflect the prediction performance of AI model directlyNot reflect the system/link performance directly | Not reflect the system/link performance directly |

Note1: The above analysis shall not give an indication about whether/which metric is supported or specified.

Note2: Monitoring performance of the above alternatives are not addressed in the table.

*Data collection:*

At UE side for UE-side AI/ML model:

- UE reporting to NW supported/preferred configurations of DL RS transmission.

- Trigger/initiating data collection considering:

- Option 1: data collection initiated/triggered by configuration from NW.

- Option 2: request from UE for data collection.

- Signalling/configuration/measurement/report for data collection, e.g., signalling aspects related to assistance information (if supported), Reference signals, content/type of the collected data, configuration related to Set A and/or Set B, information on association/mapping of Set A and Set B

- Assistance information from Network to UE for UE data collection for categorizing the data for the purpose of differentiating characteristics of the data (if supported). The assistance information should preserve privacy/proprietary information.

At NW side for NW-side AI/ML model:

- Mechanism related to the reporting.

- Additional information for content of the reporting.

- Reporting overhead reduction.

- Signalling/configuration/measurement/report for data collection, e.g., signalling aspects related to assistance information (if supported), Reference signals.

Regarding data collection for NW-side AI/ML model regarding the contents of collected data:

- Opt.1: M1 L1-RSRPs (corresponding to M1 beams) with the indication of beams (beam pairs) based on the measurement corresponding to a beam set, where M1 can be larger than 4, if applicable.

- Opt.2: M2 L1-RSRPs (corresponding to M2 beams) based on the measurement corresponding to a beam set, where M2 can be larger than 4, if applicable.

- Opt.3: M3 beam (beam pair) indices based on the measurement corresponding to a beam set, where M3 can be larger than 4, if applicable.

- Note: Overhead, UE complexity and power consumption are to be considered for the above options.

Regarding data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, the following approaches have been identified for overhead reduction:

- the omission/selection of collected data

- the compression of collected data

- Note1: For the different purposes of data collection, the overhead reduction mechanisms and corresponding specification impacts may be different.

- Note2: Support of any mechanism(s) (if necessary) for each LCM purpose and the potential spec impact (if any) are separate discussions

- Note 3: UE complexity and power consumption should be considered

Regarding data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, the following reporting signalling for beam-specific aspects maybe applicable:

- L1 signalling to report the collected data

- Higher-layer signalling to report the collected data

- At least not applicable to AI/ML model inference

- Note1: Higher layer signalling design is up to RAN2

- Note2: Whether each signalling applicable to each LCM purpose is a separate discussion

- Note3: The legacy signalling principle (e.g. RSRP reporting for L1) can be re-used

*Model Inference related*:

In order to facilitate the AI/ML model inference:

- Enhanced or new configurations/UE reporting/UE measurement, e.g., enhanced or new beam measurement and/or beam reporting

- Enhanced or new signalling for measurement configuration/triggering

- Signalling of assistance information (if applicable)

For BM-Case1 and BM-Case2 with a UE-side AI/ML model:

- Indication of the associated Set A from network to UE, e.g., association/mapping of beams within Set A and beams within Set B if applicable

- Beam indication from network for UE reception, which may or may not have additional specification impact (e.g., legacy mechanism may be reused), particularly:

- how to perform beam indication of beams in Set A not in Set B. Note: also applicable to NW-side AI/ML model. Note: At least for BM-Case1 with a UE-side AI/ML model, the legacy TCI state mechanism can be used to perform beam indication of beams

- Note: For DL beam pair prediction, there is no consensus to support the reporting of the predicted Rx beam(s) (e.g., Rx beam ID, Rx beam angle information, etc) from the UE to the network.

- Predicted L1-RSRP(s) corresponding to the DL Tx beam(s) or beam pair(s)

- Whether/how to differentiate predicted L1-RSRP and measured L1-RSRP

- Confidence/probability information related to the output of AI/ML model inference (e.g., predicted beams)

For BM-Case1 and BM-Case2 with a NW-side AI/ML model:

- L1 beam reporting enhancement for AI/ML model inference:

- UE to report the measurement results of more than 4 beams in one reporting instance

- Other L1 reporting enhancements can be considered

For BM-Case1 with a UE-side AI/ML model:

- L1 signalling to report the following information of AI/ML model inference to NW:

- The beam(s) that is based on the output of AI/ML model inference.

For BM-Case2 with a UE-side AI/ML model:

- L1 signalling to report the following information of AI/ML model inference to NW:

- The beam(s) of N future time instance(s) that is based on the output of AI/ML model inference.

- Information about the timestamp corresponding the reported beam(s).

For BM-Case 2:

- Reporting information about measurements of multiple past time instances in one reporting instance. Notes: Only applicable to NW-side AI/ML model. The potential performance gains of measurement reporting should be justified by considering UCI payload overhead.

*Assistance information:*

Regarding the explicit assistance information from UE to network for NW-side AI/ML model, RAN1 has no consensus to support the following information

- UE location

- UE moving direction

- UE Rx beam shape/direction

Regarding the explicit assistance information from network to UE for UE-side AI/ML model, RAN1 has no consensus to support the following information

- NW-side beam shape information

- E.g., 3dB beamwidth, beam boresight directions, beam shape, Tx beam angle, etc.

- Note: Other information (e.g., relative information) of Tx beam(s) preserving sensitive proprietary information is a separate discussion

- e.g., some information following the same principle of Rel-17 positioning agreement

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consistency / association of Set B beams and Set A beams across training and inference is beneficial from performance perspective.

Note: Whether specification impact is needed is a separate discussion.

### 7.1.4 Positioning accuracy enhancements

***Items considered for studying the necessity, feasibility, potential specification impact***:

*AI/ML functionality and model identification:*

- Validity conditions, e.g., applicable area/[zone/]scenario/environment and time interval, etc.

- Model capability, e.g., positioning accuracy quality and model inference latency.

- Conditions and requirements, e.g., required assistance signalling and/or reference signals configurations, dataset information.

- Note: the above-mentioned examples and terms "validity conditions", "model capability", and "Conditions and requirements" can be referred to the conditions and additional conditions discussed in the context of the model identification and functionality identification in clause 4.2.

*Training data generation* for AI/ML based positioning:

- The following options of entity and mechanisms to generate ground-truth label are identified:

- UE with estimated/known location generates ground-truth label and corresponding label quality indicator

- Based on non-NR and/or NR RAT-dependent and/or NR RAT-independent positioning methods

- At least for UE-based positioning with UE-side model (Case 1) and UE-assisted positioning with UE-side model (Case 2a)

- Network entity generates ground-truth label and corresponding label quality indicator

- Based on non-NR and/or NR RAT-dependent and/or NR RAT-independent positioning methods

- At least for UE-assisted/LMF-based positioning with LMF-side model (Case 2b), NG-RAN node assisted positioning with gNB-side model (Case 3a) and NG-RAN node assisted positioning with LMF-side model (Case 3b)

- At least PRU is identified to generate ground-truth label for UE-based positioning with UE-side model (Case 1) and UE-assisted positioning with UE-side model (Case 2a)

- At least LMF with known PRU location is identified to generate ground-truth label for UE-assisted/LMF-based positioning with LMF-side model (Case 2b) and NG-RAN node assisted positioning with LMF-side model (Case 3b)

- At least network entity with known PRU location is identified to generate ground-truth label for NG-RAN node assisted positioning with gNB-side model (Case 3a)

- Note: User data privacy needs to be preserved

- The following options of entity to generate other training data (at least measurement corresponding to model input) are identified:

- For UE-based with UE-side model (Case 1) and UE-assisted positioning with UE-side (Case 2a) or LMF-side model (Case 2b)

- PRU

- UE

- For NG-RAN node assisted positioning with Network-side model (Case 3a and Case 3b)

- TRP

- Note: Transfer of training data from the entity generating training data to a different entity is not precluded and associated potential specification impact is to be considered

*Data collection* for AI/ML based positioning:

Regarding data collection for AI/ML based positioning, at least the following information of data with potential specification impact are identified.

- Ground-truth label

- Report from the label data generation entity

- Measurement (corresponding to model input)

- Report from the measurement data generation entity

- Quality indicator

- For and/or associated with ground-truth label and/or measurement

- Report from the label and/or the measurement data generation entity and/or as request from a different (e.g., data collection, etc.) entity

- RS configuration(s)

- At least for deriving measurement

- Request from data generation entity (UE/PRU/TRP) to LMF and/or as LMF assistance signalling to UE/PRU/TRP

- Note 1: There may not be any enhancements on top of existing RS configuration(s) or any new RS configuration(s) for positioning measurement

- Time stamp

- At least for and/or associated with collected data

- Separate time stamp for measurement and ground-truth label, when measurement and ground-truth label are generated by different entities

- Report From data generation entity together with collected data and/or as LMF assistance signalling

- Note 2: There may not be any enhancements on top of time stamp in existing positioning measurement report or any new time stamp report for positioning measurement

- Note 3: Whether and how the above information can be applied to different aspects of AI/ML LCM (e.g., training, updating, monitoring, etc.) can be discussed

- Note 4: Transfer of data from the entity generating data to a different entity is not precluded from RAN1 perspective

- Note 5: If any specification impact is identified, the impact may be different between positioning use cases (Case 1/2a/2b/3a/3b).

- Note 6: The necessity of other information (e.g., scenario identifier. LOS/NLOS condition, timing error, etc.) for data collection can be discussed

- Details of request/report of label and/or other training data, and to enable delivering the collected label and/or other training data to the training entity when the training entity is not the same entity to obtain label and/or other training data

- Assistance signalling indicating reference signal configuration(s) to derive label and/or other training data

- Request/report of training data: Ground-truth label; Measurement corresponding to model input; Associated information of ground-truth label and/or measurement corresponding to model input

- Assistance signalling and procedure to facilitate generating training data: Reference signal (e.g., PRS/SRS) configuration(s) and configuration identifier; Assistance information, e.g., between LMF and UE/PRU, for label calculation/generation, and label validity/quality condition, etc.

- Note: whether such assistance signalling and procedure can be applied to other aspect(s) of AI/ML model LCM can also be discussed

- Notes: Study may consider different entity to generate training data as well as different types of training data when applicable. Study considers both of the following cases when applicable: when the training entity is the same entity to generate training data, and when the training entity is not the same entity to generate training data

*Model monitoring:*

- Assistance signalling and procedure at least for UE-side model

- Report/feedback and procedure at least for Network-side model

- Note: study is applicable to both of the following cases:

- Model inference and model monitoring at the same entity

- Entity to perform the model monitoring is not the same entity for model inference

- Data for computing monitoring metric:

- If monitoring based on model output: e.g., estimated UE location corresponding to model output for direct AI/ML positioning, estimated intermediate parameter(s) corresponding to model output for AI/ML assisted positioning, ground-truth label corresponding to model inference output for both direct and AI/ML assisted positioning

- If monitoring based on model input: e.g., measurement corresponding to model inference input.

- Assistance signalling from LMF to UE/PRU/gNB for UE/gNB-side model monitoring.

- Assistance signalling from UE/PRU for NW-side model monitoring.

- If certain type of data is necessary for computing monitoring metric:

- How an entity can be used to provide the given type of data for calculating monitoring metric: companies requested to report their assumption of the entity (or entities) used to provide the given type of data for calculating monitoring metric for each case

- Potential signalling for provisioning of the given type of data for calculating associated monitoring metric

- Potential assistance signalling and procedure to facilitate an entity providing data for calculating monitoring metric

- Potential UE-network interaction: e.g., model monitoring decision indication between UE and network

- Entity to derive monitoring metric

- UE at least for Case 1 and 2a (with UE-side model)

- gNB at least for Case 3a (with gNB-side model)

- LMF at least for Case 2b and 3b (with LMF-side model)

- For AI/ML based positioning, LMF for Case 2a (with UE-side model) and Case 3a (with gNB-side model) is identified as the entity to derive the monitoring metric at least when monitoring is based on provided ground-truth label (or its approximation).

- If model monitoring does not require ground-truth label (or its approximation).

- Statistics of measurement(s) compared to the statistics associated with the training data. Note: the measurement(s) may or may not be the same as model input.

- Examples used in contributions: norm of model input, mean, min/max of some statistics related to measurement and/or model input, median or data temporal/spatial distribution

- Statistics of model output compared to the statistics associated with the training data and/or its own previous inference output

- Examples used in contributions: mean, standard deviation, variance, etc. of some statistics related to model output

- For monitoring UE-side and gNB-side model for AI/ML based positioning:

- Signalling from LMF to facilitate the monitoring entity to derive the monitoring metric (if needed)

- Signalling from monitoring entity to request measurement(s) (if needed)

- Signalling for potential request/report of monitoring metric (if needed)

- Note: there may not be any specification impact

- For monitoring LMF-side model for AI/ML based positioning

- Signalling from LMF to request measurement(s) (if needed)

- Assistance signalling and procedure, e.g., RS configuration(s) for measurement, measurement statistics as compared to the model input statistics of the training data, etc.

- Report of the calculated metric and/or model monitoring decision

- If model monitoring requires and is provided ground-truth label (or its approximation)

- Monitoring metric: statistics of the difference between model output and provided ground-truth label.

- Examples used in contributions: mean, standard deviation, instantaneous value, threshold of ground-truth label (or its approximation)

- For monitoring UE-side and gNB-side model for AI/ML based positioning:

- Signalling from monitoring entity to request ground-truth label (if needed)

- Signalling from monitoring entity to request model output (if needed)

- Signalling for potential request/report of monitoring metric (if needed)

- For monitoring LMF-side model for AI/ML based positioning

- Signalling from LMF to request measurement(s) (if needed)

- Provisioning of ground-truth label and associated label quality.

- Assistance signalling and procedure, e.g., from LMF to UE/gNB indicating ground-truth label and/or measurement, etc.

- Report of the calculated metric and/or model monitoring decision

- Note: No extensive evaluation results on model monitoring metric comparison have been carried out

- Note: There is no consensus during SI on whether monitoring metric will have spec impact

*Model Inference related:*

- For direct AI/ML positioning (Case 2b and 3b), type of measurement(s) as model inference input considering performance impact and associated signalling overhead

- Potential new measurement: CIR/PDP

- Existing measurement: e.g., RSRP/RSRPP/RSTD

- Note: Details of potential new measurement and/or potential enhancement to existing measurement is to be studied.

- For AI/ML assisted positioning with UE-assisted (Case 2a) and NG-RAN node assisted positioning (Case 3a):

- Measurement report to carry model output to LMF

- New measurement report: e.g., ToA, path phase

- Existing measurement report: e.g., RSTD, LOS/NLOS indicator, RSRPP

- Enhancement of existing measurement report: e.g., soft information/high resolution of RSTD

- At least the following types of model inference output are identified as candidates providing performance benefits:

- Timing estimation

- Note: the report to LMF is derived based on and maybe different from the model inference output

- LOS/NLOS indicator

- Assistance signalling and procedure to facilitate model inference for both UE-side and Network-side model

- RS configurations

The specification impact related to the following items is assessed:

- Types of measurement as model inference input

- new measurement

- existing measurement

- UE is assumed to perform measurement as model inference input for Case 1, Case 2a and Case 2b; TRP is assumed to perform measurement as model inference input for Case 3a and Case 3b

- Report of measurements as model inference input to LMF for LMF-side model (Case 2b and Case 3b)

- For AI/ML assisted positioning, new measurement report and/or potential enhancement of existing measurement report as model output to LMF for UE-assisted (Case 2a) and NG-RAN node assisted positioning (Case 3a)

- Assistance signalling and procedure to facilitate model inference for both UE-side and Network-side model

- New and/or enhancement to existing assistance signalling

- Note: Whether such assistance signalling and procedure can be applied to other aspect(s) of AI/ML model LCM can also be discussed

For direct AI/ML positioning with LMF-side model (Case 2b and 3b), the following types of measurement report areidentified if beneficial and necessary (e.g., tradeoff positioning accuracy requirement and signalling overhead),

- Take into account that existing Rel-16/17 measurement and/or expected Rel-18 measurement report may contain timing, power and phase information of the channel response

- measurement report, which contains timing, power and phase information of the channel response

- At least for Case 3b

- Measurement report, which contains timing and power information of the channel response

- Measurement report, which contains timing information of the channel response

- Note: Combinations of multiple measurement reports and/or post processing of the measurement reports are not precluded

For direct AI/ML positioning with LMF-side model (Case 2b and 3b), the following types of measurement report with potential specification impact have been studied for AI/ML based positioning accuracy enhancement

- Measurement report, which contains timing, power and phase information of the channel response

- If support, potential specification impact including new measurement report or enhancement to existing measurement report

- E.g., truncation, [feature extraction,] alignment of sample/path determination

- Measurement report, which contains timing and power information of the channel response

- If support, potential specification impact including new measurement report or enhancement to existing measurement report

- E.g., truncation, [feature extraction,] alignment of sample/path determination

- Measurement report, which contains timing information of the channel response

- If support, potential specification impact including enhancement to existing measurement report

- E.g., alignment of sample/path determination

*LCM:*

- For AI/ML based positioning accuracy enhancement, at least for Case 1 and Case 2a (model is at UE-side)

- which aspects should be specified as conditions of a Feature/FG available for functionality-based LCM.

- which aspects should be considered as additional conditions, and how to include them into model description information during model identification for model ID-based LCM.

## 7.2 Protocol aspects

In this clause, considering the use cases and as per RAN1 input, aspects related to life cycle management signalling procedures, model identification, data collection, model transfer/delivery, UE capability reporting, and applicability-related reporting are studied.

### 7.2.1 Common framework

#### 7.2.1.1 Signalling procedures for model and functionality life cycle management

As per the functional framework in Figure 4.4-1, in this clause the signalling procedures for different scenarios for model-ID-based management and/or functionality-based management are exemplified. The procedures can at least be considered for UE-side models. From clause 4.2, these can include scenarios for which the management decision is taken by the network or by the UE. For network-side decision, this can be either network-initiated, or UE-initiated and requested to the network. While for UE-side decision, this can be either event-triggered as configured by the network and where the UE’s decision is reported to the network, or UE-autonomous, with or without UE’s decision being reported to the network.

Note: The mapping of these scenarios to specific use cases can be left to RAN1.

Note: The scenarios discussed below shall not imply support for all potential functionality and/or model Management Instructions (e.g., (de)activation, selection, switching, fallback, etc.) for every use case.

Note: In the figures below, Management Request/Management Instruction/Management Decision Report may include details about the model/functionality selection, (de)activation, switching or fallback.

**- Decision by the network**

**o Network-initiated**



Figure 7.2.1.1-1: Network decision, network-initiated AI/ML management

The case where the LCM decision is taken and initiated by the network is depicted in Figure 7.2.1.1-1.

Note: The Management Instruction may be a result of model/functionality performance monitoring at the network.

Note: The Management Instruction may include information about the model or functionality.

**o UE-initiated and requested to the network**



Figure 7.2.1.1-2: Network decision, UE-initiated AI/ML management

The case where the LCM decision is taken by the network but where the request is initiated by the UE is depicted in Figure 7.2.1.1-2.

Note: The Management Request may be a result of model/functionality monitoring at the UE.

Note: In response to the Management Request, the network may send a Management Instruction to the UE.

Note: The Management Request may include information about the model or functionality.

Note: The network may accept or reject the Management Request from the UE.

Note: The Management Request may include information related to model/functionality performance metrics.

Note: The Management Instruction may include information about the model or functionality.

**- Decision by the UE**

**o Event-triggered as configured by the network, UE’s decision is reported to the network**



Figure 7.2.1.1-3: UE decision, event-triggered as configured by the network

The case where the LCM decision is taken by the UE according to prior network configuration is depicted in Figure 7.2.1.1-3.

Note: Use case-specific events/conditions may be configured by the network for event-triggered AI/ML management at the UE.

Note: UE may send a Management Decision Report to the network following event-triggered AI/ML management at the UE.

Note: The Management Decision Report may include information about the model or functionality.

**o UE-autonomous, UE’s decision is reported to the network**



Figure 7.2.1.1-4: UE autonomous, decision reported to the network

The case where the LCM decision can autonomously be taken by the UE is depicted in Figure 7.2.1.1-4.

Note: The UE may be configured to send a Management Decision Report to the network upon performing a model/functionality Management Decision.

**o UE-autonomous, UE’s decision is not reported to the network**

For the case where the LCM decision can autonomously be taken by the UE and where the decision is not reported to the network, the AI/ML management is transparent from a network perspective.

#### 7.2.1.2 Model identification and meta information

According to the functional framework in Figure 4.4-1, a model ID can be used within functions and for different data/information/instruction flows to identify an AI/ML model. For example, a model ID could eventually be associated to the selection/(de)activation/switching of a model or linked to the "Model Transfer/Delivery" information.

RAN2 assumes that a model ID can be globally unique, e.g., allowing for proper model validation and model testing procedures.

Note: How to ensure model ID uniqueness is out of RAN2 scope.

Note: Details of model training, validation and testing are out of RAN2 scope.

Additionally, to manage or control AI/ML models, some meta information about the models may be needed.

Note: Details on the relationship between model IDs and meta information for purposes of model control and management can be addressed during a normative phase.

#### 7.2.1.3 Data collection

Data collection plays a crucial role in enabling the different use cases. Therefore, it is important to define the best approaches for collecting data to support UE-side and network-side model inference, monitoring, and training.

Table 7.3.1.2-1 lists existing data collection mechanisms available in current RAN specifications for the UE to report measurements to another entity acting as termination point for this data. As highlighted in clause 4.2, the analysis/selection of the data collection frameworks should focus on the RRC CONNECTED state for both data generation and reporting. As such, the Table can provide useful insights into existing methods with respect to various categories identified as relevant for data collection method selection.

Table 7.3.1.2-1. Existing data collection methods identified.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Involved network entity (termination point)** | **RRC state to generate data** | **Max payload size per reporting\*** | **Contents to be collected** | 1) **End-to-End report latency\*\*** | **Report type** | **Security and Privacy** |
| **Method: Logged MDT** |
| TCE/OAM(Data can be utilized by gNB) | IDLE / INACTIVE | <9kbyte | - L3 cell/beam measurements- location information- sensor information- timing information | 1) Procedure latency\*\*\*:- Latency to enter CONNECTED state- Latency to receive gNB request signalling (~20ms)2) Air interface signalling latency\*\*\*\*: - ~20ms (RRC)3) Other latency:- Forwarding latency between gNB and TCE | Upon gNB request after entering RRC\_CONNECTED | AS security via RRC messagePrivacy via user consent  |
| **Method: Immediate MDT** |
| TCE/OAM(Data can be utilized by gNB) | CONNECTED | <9kbyte | - L3 cell/beam measurements- location information- sensor information | 1) Procedure latency:- Report interval:  120ms~30min for periodic report TTT for event triggered report2) Air interface signalling latency:- ~20ms (RRC)3) Other latency:- Forwarding latency between gNB and TCE  | - Event triggered- Periodic reporting  | AS security via RRC messagePrivacy via user consent |
| **Method: L3 measurements** |
| gNB | CONNECTED | <9kbyte | L3 cell/beam measurements | 1) Procedure latency:- Report interval:  l20ms~30min for periodic report TTT for event triggered report2) Air interface signalling latency:- 20ms (RRC) | - Event triggered report- Periodic reporting | AS security via RRC message |
| **Method: L1 measurement (CSI reporting)** |
| gNB | CONNECTED | <1706bit in PUCCH<3840bit in PUSCH | L1 CSI measurement | 1) Procedure latency:- Report interval:  4-320 slot for periodic and semi-persistent report  0-32 slot after reception of DCI for aperiodic report 2) Air interface signalling latency:- 1 TTI (PUCCH)  | - Aperiodic report- Semi-persistent report- Periodic report | No AS security |
| **Method: UE Assistance Information (UAI)** |
| gNB | CONNECTED | <9kbyte | Assistance information to show UE preference | 1) Procedure latency:- Upon generation of UE's preference2) Air interface signalling latency:- ~20ms (RRC) | Up to UE implementation when to report | AS security via RRC message |
| **Method: Early measurements** |
| gNB | IDLE / INACTIVE | <9kbyte | L3 cell/beam measurements | 1) Procedure latency:- Latency to enter CONNECTED state- Latency to receive gNB request signalling (~20ms)2) Air interface signalling latency: - ~20ms (RRC) | Upon gNB request after entering RRC\_CONNECTED | AS security via RRC message |
| **Method: LPP** |
| LMF | CONNECTED | <9kbyte | Location information | 1) Procedure latency:- Latency to get upper layer trigger (for UE triggered)- Or latency to receive network request message (~20ms)2) Air interface signalling latency: - ~20ms (RRC)3) Other latency:- Forwarding latency between gNB and LMF | - UE-triggered- Network-triggered | AS security via RRC message |
| \*: The payload size doesn't consider signalling overhead.\*\*: The End-to-End report latency is the latency from availability of the measurement report at the UE side to the availability of the measurement report at the terminated network entity. The time to generate data or perform measurements depends on RAN1/RAN4 specification.\*\*\*: Procedure latency is the latency caused by procedures, including procedure to ready for reporting (e.g., entering CONNECTED state, report interval).\*\*\*\*: Air interface signalling latency is the latency to transmit one report, e.g., RRC signalling latency or PUCCH signalling latency. |

##### 7.2.1.3.1 Considerations for network-side data collection

A set of general data collection principles is expected to be considered for network-side model training. These include:

- UE to support data logging,

- UE to report the collected data periodically, event-based, and on-demand,

- The UE memory, processing power, energy consumption, signalling overhead should be considered.

Note: The above principles can be revised depending on RAN1 requirements.

Furthermore, and regarding the use cases in this study, the following is considered.

For CSI and beam management use cases, the training of network-side models can consider both gNB and OAM-centric data collection mechanisms. The gNB-centric data collection implies that the gNB can configure the UE to initiate/terminate the data collection procedure. The potential impact of L3 signalling for the reporting of collected data should be assessed.

On the other hand, OAM-centric data collection implies that the OAM provides the configuration (via the gNB) needed for the UE to initiate/terminate the data collection procedure. MDT framework can be considered to achieve this. The potential impact on MDT for RRC\_CONNECTED state should be assessed.

For positioning use cases, when considering LMF-side inference, it is assumed that the LPP protocol should be applied to the data collected by UE and terminated at LMF, while the NRPPa protocol should be applied to the data collected by gNB and terminated at LMF. While for LMF-side performance monitoring, it is assumed that the LPP protocol should be applied to the data collected by UE and terminated at LMF, while the NRPPa protocol should be applied to the data collected by gNB and terminated at LMF.

Note: For gNB- and OAM-centric data collection, there may be a need to consult with RAN3 and SA5 whether/how OAM is to be involved.

Note: For possible impacts due to positioning use cases, there may be a need to consult with RAN3 whether/how NRPPa is to be involved.

##### 7.2.1.3.2 Data collection for UE-side model training

The following proposals were discussed in RAN2:

1. UE collects and directly transfers training data to the Over-The-Top (OTT) server;

1a) OTT (TRansparent)

1b) OTT (non-TRansparent)

2. UE collects training data and transfers it to Core Network. Core Network transfers the training data to the OTT server.

3. UE collects training data and transfers it to OAM. OAM transfers the needed data to the OTT server.

RAN2 did not study or analyse these proposals and did not agree to requirements or recommendations.

#### 7.2.1.4 Model transfer/delivery

Whether there is a need to consider standardised solutions for transferring/delivering AI/ML model(s) is unclear from the outcome of the present study. Nonetheless, to support AI/ML model transfer/delivery, the following solutions are considered:

- Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

- Solution 2a: Core Network (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.

- Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.

- Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.

- Solution 2b: Core Network (except LMF) can transfer/deliver AI/ML model(s) to UE via User Plane (UP) data.

- Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.

- Solution 4a: OTT server can transfer/deliver AI/ML model(s) to UE (e.g., transparent to 3GPP).

- Solution 4b: OAM can transfer/deliver AI/ML model(s) to UE.

Note: The relationships between model transfer/delivery solutions and use cases can be derived from what is captured in clauses 7.3.2, 7.3.3, and 7.3.4.

The following areas are considered to evaluate the different model transfer/delivery solutions:

- A1: Large, no upper limit model/model parameter size,

- A2: Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs),

- A3: Network controllability on model transfer/delivery (e.g., management decision at gNB),

- A4: Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB).

For every model transfer/delivery solution, each of the above areas is analysed, focusing on the current status and gaps, and the potential impacts on RAN specification. The analysis is shown in the Tables below.

Table 7.3.1.4-1 Analysis of current status and gaps, and potential
RAN specification impact for Solution 1a

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact** |
| A1. Large, no upper limit model/model parameter size | Model size >45kBytes is not supported based on existing number of RRC segments | Extension of the number of RRC segments is required to support models larger than 45kBytes |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | Transmission is restarted upon mobility | - Requires service continuity support for SRBs with segmentations.- Xn/NGAP enhancement(s) for model transfer/delivery continuity |
| A3. Network controllability on model transfer/delivery and management at gNB | Management and interaction between UE and gNB is not supported | Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion indication, etc.) when model management at gNB |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | Procedure latency depends on model size and SRB priority | Impact on SRB in DL, e.g., a new SRB with configurable priority, etc. |

Table 7.3.1.4-2 Analysis of current status and gaps, and potential
RAN specification impact for Solutions 2a and 3a

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact** |
| A1. Large, no upper limit model/model parameter size | - Model size >45kBytes is not supported based on existing number of RRC segments- Core Network supports NAS signalling segmentation- LMF supports LPP signalling segmentation | If NAS/LMF does not do segmentation for model transfer/delivery, it may need RRC segmentation, and extension of the number of RRC segments is required to support models larger than 45kBytes |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | Supported with limitation:- For Solution 2a, support within AMF coverage area based on NAS signalling segmentation- For Solution 3a, support within LMF coverage area based on LPP signalling segmentation | Note: Supporting service continuity across AMF/LMF is out of RAN scope and needs coordination with Core Network groups |
| A3. Network controllability on model transfer/delivery and management at gNB | - For Solution 2a, gNB cannot perform management directly, considering model transfer is transparent to gNB- Management and interaction between UE and gNB is not supported | - Requires management and model transfer interaction between Core Network/LMF and gNB, e.g., via NAS signalling or NRPPa signalling when model management at gNB- Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion indication, etc.) when model management at gNB |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | Procedure latency depends on model size and SRB priority; other latency includes forwarding NAS message latency from Core Network to gNB | Impact on SRB in DL, e.g., a new SRB with configurable priority, etc. |

Note: NAS and LMF upper limits and potential impacts to NAS and LPP specifications have not been studied and feasibility on filling gaps is unknown.

Table 7.3.1.4-3 Analysis of current status and gaps, and potential
RAN specification impact for Solutions 1b

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact** |
| A1. Large, no upper limit model/model parameter size | - No model size limitation- PDU session termination at gNB is not supported | Requires PDU session termination at gNB if needed |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | Model transfer continuity if PDU session terminated at gNB is not studied | - Identify a solution to support service continuity support between gNBs when PDU session is terminated at gNB if needed- Xn/NGAP enhancement(s) for model transfer/delivery continuity |
| A3. Network controllability on model transfer/delivery and management at gNB | Management and interaction between UE and gNB appear to be feasible but not supported | Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion indication, etc.) when model management at gNB |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | - Procedure latency depends on model size, QoS requirement and DRB priority- QoS management at gNB if PDU session is terminated at gNB is not supported | Identify a solution to support QoS management at gNB for model transfer when PDU session is terminated at gNB if needed |

Table 7.3.1.4-4 Analysis of current status and gaps, and potential
RAN specification impact for Solutions 2b and 3b

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact** |
| A1. Large, no upper limit model/model parameter size | No model size limitation | - No RAN impact- Note: The detail procedure of model transfer from Core Network/LMF to UE is out of RAN scope |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | - For Solution 2b, supported- For Solution 3b, depends on Rel-18 CT1 solution LPP message over a user plane connection between UE and LMF | Note: supporting service continuity across LMF is out of RAN scope |
| A3. Network controllability on model transfer/delivery and management at gNB | gNB cannot perform model management directly | Requires management and model transfer interaction between Core Network/LMF and gNB when model management at gNB |
| Management and interaction between UE and gNB is not supported | Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion, etc.) when model management at gNB |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | - Procedure latency depends on model size, QoS requirement and DRB priority- Other latency includes forwarding data from Core Network to gNB | Note: The detail QoS requirement on Core Network for model transfer/delivery is out of RAN scope |

Table 7.3.1.4-5 Analysis of current status and gaps, and potential
RAN specification impact for Solutions 4a

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact** |
| A1. Large, no upper limit model/model parameter size | No model size limitation | No RAN impact |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | - If model transfer/delivery from OTT server via Core Network, supported- If model transfer/delivery from OTT server via LMF, depends on Rel-18 CT1 solution LPP message over a User Plane connection between UE and LMF | Note: supporting service continuity across LMF is out of RAN scope |
| A3. Network controllability on model transfer/delivery and management at gNB | Model transfer/delivery is transparent to RAN | - Requires management and model transfer interaction between OTT server and gNB when model management at gNB- Note: it is unclear whether this is within RAN scope- Requires interaction between UE and gNB for the network controllability of the model transfer/delivery (e.g., model identification, model transfer completion, etc.) if management is in gNB |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | - Procedure latency depends on model size, QoS requirement and DRB priority- Other latency includes forwarding data from OTT server to gNB | Note: The detail QoS requirement for model transfer/delivery of solution 4a is out of RAN scope |

Table 7.3.1.4-6 Analysis of current status and gaps, and potential
RAN specification impact for Solutions 4b

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Current status and Gaps** | **Potential RAN specification impact(Note: whether and how to support model transfer/delivery from OAM to gNB and OAM to UE directly is out of RAN scope)** |
| A1. Large, no upper limit model/model parameter size | - Over Control Plane (CP) signalling: model size >45kBytes is not supported based on existing number of RRC segments if OAM does not do segmentation for model transfer/delivery- Over, e.g., IP: no model size limitation, but direct connection between OAM and UE is not supported | - Over Control Plane (CP) signalling: If OAM does not do segmentation for model transfer/delivery, it may need RRC segmentation, and extend RRC segment number if model size larger than 45kBytes- Over, e.g., IP: NOTE: whether and how to support direct connection between OAM and UE is out of RAN scope |
| A2. Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs) | Support within OAM coverage |  |
| A3. Network controllability on model transfer/delivery and management at gNB | gNB cannot perform model management directly | Note: support management and model transfer interaction between OAM and gNB is out of RAN scope |
| A4. Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB) | - Over Control Plane (CP) signalling:1) Procedure latency depends on model size and SRB priority2) other latency includes forwarding data from OAM to gNB- Over, e.g., IP: direct connection between OAM and UE is not supported | - Over Control Plane (CP) signalling:Note: The detail QoS requirement for model transfer/delivery of solution 4b is out of RAN scope- Over, e.g., IP:Note: whether and how to support latency, QoS requirement between OAM and UE is out of RAN scope |

Note: For Solution 4b, RAN2 discussed the following two solutions but did not study or analyse their feasibility:
- OAM can transfer/deliver AI/ML models to UE via OAM→RAN→UE, where Control Plane (CP) signalling is used for RAN→UE.
- OAM can transfer/deliver AI/ML models to UE via OAM→UE, e.g., via IP tunnel.

A reactive and a proactive approach for initiating a model transfer/delivery can be considered in a normative phase. For the reactive approach, an AI/ML model is transferred/delivered (i.e., downloaded) to the UE when needed. This could typically happen due to changes in scenarios, configurations, sites, etc. While for the proactive model transfer/delivery approach, an AI/ML model is pre-download to the UE, and a model switch can typically be performed due to changes in scenarios, configurations, sites, etc.

#### 7.2.1.5 UE capability reporting

The legacy UE capability framework serves as the baseline to report UE’s supported AI/ML-enabled Feature/FG. Therefore, for CSI and beam management use cases, this information is indicated in UE AS capability in RRC (e.g., *UECapabilityEnquiry/UECapabilityInformation*). While for positioning use cases, it is indicated by the positioning capability as defined in LPP.

Further discussions concerning UE capability details (e.g., granularity of Feature/FG, content, structure of the related UE capabilities, etc.) can be carried during a normative phase.

#### 7.2.1.6 Reporting applicability-related information

AI/ML models for a given use case may be tailored towards and applicable to specific scenarios, locations, configuration, deployments, among other factors. In this regard, it is acknowledged that AI/ML models may undergo updates, such as model changes, as an inherent part of their development. Therefore, to ensure efficient network control and management, especially associated to what concerns the UE-side, UEs might have the ability to indicate relevant information about their supported AI/ML models and concerning AI/ML functionalities to the network. This can allow the network to perform decisions regarding, e.g., the (de)activation, or switching of AI/ML functionalities and AI/ML models.

The previously mentioned information could in principle be understood as "applicability-related information" in which the UE could, for example, report to the network conditions under which a model/functionality is applicable/suitable, or whether model(s)/functionality(es) are (non)applicable under the current context. Note, however, that the existing UE capability reporting framework cannot be used for such purposes.

Note: How and whether there is a need to enable UEs to report applicability-related information can be further discussed and defined in a normative phase. Mechanisms such as UE Assistance Information can eventually be used as example.

Two UE reporting types are identified to convey this additional information:

- *"reactive"* reporting, and

- *"proactive"* reporting.

A reactive reporting would involve the UE to provide information to the network upon receiving an action from it.

While a proactive reporting would involve the UE to provide information to the network without necessarily receiving an action from it. For example, the UE might proactively inform the RAN of updates/changes to its supported model(s) or functionality(es).

Note: Whether necessary signalling from network is needed for proactive UE reporting can be discussed in a normative phase.

Note: Whether there is a need for the network to report to the UE applicability-related information of AI/ML models and/or AI/ML functionalities can be discussed in a normative phase.

### 7.2.2 CSI feedback enhancement

The following set of objectives have been identified for the two-sided CSI compression use case. Firstly, to ensure that the UE part and network part of the models are configured and applied according to their applicable scenarios and configuration. Secondly, to ensure that models match properly, ensuring that the CSI generation part used at the UE corresponds to the CSI reconstruction part employed at the gNB. Thirdly, to allow for seamless operation, requiring the simultaneous (de)activation and switching of the two-sided model.

Regarding the last point above, for the two-sided model CSI compression use cases, the selection, (de)activation, switching, and fallback of AI/ML models or AI/ML functionalities can be initiated by either the UE or the gNB. For which it is important to distinguish the various cases and understand their applicability to UE-side versus network-side models.

For data collection, model transfer/delivery, and function-to-entity mapping analysis, various scenarios unfold for both the two-sided CSI compression use case, as well as for the UE-side CSI prediction use case, when the data generation and termination entities differ. For instance, for:

- Model Training:

o For the two-sided CSI compression use case, training data can be generated by either the UE or the gNB, depending on specific requirements, while the termination point for training data may include the gNB, OAM, Over-The-Top (OTT) server or UE.

 Note: RAN2 identified the case in which Core Network may be used for model training. However, no study was conducted since this is beyond the scope of this Working Group.

o For the UE-side CSI prediction use case, training data can be generated by the UE, while the termination point for training data may include the UE or a UE-side OTT server.

 Note: RAN2 identified the cases in which OAM or Core Network may be used for UE-side model training. However, no study was conducted since this is beyond the scope of this Working Group.

 Note: RAN2 identified the case in which gNB may be used for UE-side model training. However, no conclusion was reached, as this depends on the RAN1 progress.

- Inference:

o For the two-side CSI compression use case:

 For network part of two-sided model inference, the UE can generate the necessary input data while the termination point for this input data lies within the gNB, where the inference process is performed.

 For UE part of two-sided model inference, input data is internally available at UE, where the inference process is performed.

o For the UE-side CSI prediction use case:

 For UE-side model inference, input data is internally available at UE, where the inference process is performed.

- Management:

o For the two-sided CSI compression use case, the model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) is performed by the gNB.

 Note: RAN2 identified the case in which the control is performed by the UE. However, no conclusion was reached, as this depends on the RAN1 progress.

o For the UE-side CSI prediction use case:

 The model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the UE when the monitoring resides within the UE.

 The model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the gNB when the monitoring resides within the gNB or UE.

o Monitoring:

 The UE monitors the performance of its UE-side model.

 For monitoring at the network side of UE-side model, the UE can generate, if needed, calculated performance metrics or data required for performance metric calculation, while the termination point for these is the gNB.

### 7.2.3 Beam management

For beam management, the selection, (de)activation, switching, and fallback of models or functionalities can also be initiated by either the UE or the gNB. For which it is important to distinguish the various cases and understand their applicability to UE-side versus network-side models.

For data collection, model transfer/delivery, and function-to-entity mapping analysis, various scenarios unfold when the data generation and termination entities differ. For instance, for:

- Model Training:

o For UE-side models, training data can be generated by the UE, while the termination point for training data may include the UE or a UE-side OTT server.

 Note: RAN2 identified the cases in which OAM or Core Network may be used for UE-side model training. However, no study was conducted since this is beyond the scope of this Working Group.

 Note: RAN2 identified the case in which gNB may be used for UE-side model training. However, no conclusion was reached, as this depends on the RAN1 progress.

o For gNB-side models, training data can be generated by the gNB or UE, while the termination point for training data may include the gNB, or OAM.

 Note: RAN2 identified the case in which OTT server and Core Network may be used for gNB-side model training. However, no study was conducted since this is beyond the scope of this Working Group.

- Inference:

o For UE-side model inference, input data is internally available at UE, where the inference process is performed.

o For network-side model inference, the UE can generate the necessary input data while the termination point for this input data lies within the gNB, where the inference process is performed.

- Management:

o For UE-side model, the model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the UE when the monitoring resides within the UE.

o For UE-side model, the model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the gNB when the monitoring resides within the gNB or UE.

o Monitoring:

 The UE monitors the performance of its UE-side model.

 For monitoring at the network side of UE-side model, the UE can generate, if needed, calculated performance metrics or data required for performance metric calculation, while the termination point for these is the gNB.

 For network-side model, the monitoring resides within the gNB.

### 7.2.4 Positioning accuracy enhancements

For the positioning use cases, the selection, (de)activation, switching, and fallback of models or functionalities can be initiated by either the UE, the gNB, or the LMF. For which it is important to distinguish the various cases and understand their applicability to UE-side versus network-side models.

For data collection, model transfer/delivery, and function-to-entity mapping analysis, various scenarios unfold when the data generation and termination entities differ. For instance, for:

- Model Training:

o For UE-side models, training data can be generated by the UE, while the termination point for training data may include the UE or a UE-side OTT server.

 Note: RAN2 identified the cases in which OAM or Core Network may be used for UE-side model training. However, no study was conducted since this is beyond the scope of this Working Group.

 Note: RAN2 identified the case in which LMF may be used for UE-side model training. However, no conclusion was reached, as this depends on the RAN1 progress.

o For gNB-side model, training data can be generated by the gNB, while the termination point for training data may include the gNB, or OAM.

 Note: RAN2 identified the case in which LMF may be used for gNB-side model training. However, no conclusion was reached, as this depends on the RAN1 progress.

o For LMF-side model, the LMF is the termination point for training data.

- Inference:

o For UE-side model inference, input data is internally available at UE, where the inference process is performed.

o For gNB-side model inference, input data is internally available at gNB. For this case, the UE can also generate the necessary input data while the termination point for this input data lies within the gNB where the inference process is performed.

o For LMF-side model inference, the UE or gNB can generate the necessary input data while the termination point for this input data lies within the LMF where the inference process is performed.

- Management:

o For UE-side model, the model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the UE when the monitoring resides within the UE.

o For gNB-side model, the model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) is performed by the gNB.

o The model/functionality control (e.g., selection, (de)activation, switching, fallback, etc.) may be performed by the LMF when the monitoring resides within the LMF or UE.

o Monitoring:

 The UE monitors the performance of its UE-side model.

 For monitoring at the gNB side, and if needed, calculated performance metrics or data required for performance metric calculation, can at least be generated by the gNB.

 For monitoring at the LMF side, the gNB or UE can generate, if needed, calculated performance metrics or data required for performance metric calculation, while the termination points for these metrics is the LMF.

## 7.3 Interoperability and testability aspects

### 7.3.1 Introduction

In this section, the study of requirements and testing frameworks to validate AI/ML based performance enhancements and ensuring that UE and gNB with AI/ML meet or exceed the existing minimum requirements, if applicable, are documented.

The need and implications for AI/ML processing capabilities definition is considered.

### 7.3.2 Common framework

#### 7.3.2.1 General

The general requirements and testing frameworks for AI/ML based performance enhancements mainly focus on

- how to define requirements and tests for inference

- evaluate feasibility and necessity of requirements/tests for LCM

- requirements for data collection (in particular for training) could/need be defined

Requirements/tests for training will not be studied unless training procedures are defined. The design of test should ensure performance is guaranteed and avoid that a UE can pass the test but perform poorly in the field.

For testing goals, Option 1 and/or Option 2 below will be selected depending on the test

- Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.

- FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)

- FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)

- Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.

- FFS how to define a static scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)

- FFS whether and how to define non-static specific scenarios/configurations

#### 7.3.2.2 Principles on the definition of requirements

For the definition of AI/ML requirements, the following cases related to legacy performance should be considered

- For the cases with the existing legacy performance

- Take the legacy performance as baseline for existing use cases/procedures/functionalities /measurements that are to be enhanced by AI/ML based methods

- Further study may be needed on what is baseline performance in conditions different to the requirement condition but within the expected range of operation.

- New or enhanced performance requirements/tests could be considered for existing use cases/procedures/functionalities/measurements that are to be enhanced by AI/ML based methods

- For the cases without the existing legacy performance

- New performance requirements/tests could be considered for the use cases/procedures/functionalities/measurements that are carried out or are to be enhanced by AI/ML based methods

The following procedure can be considered for defining core requirements:

- Performance monitoring procedure, including performance evaluation and decision-making procedure for AI/ML functionalities/models

- Functionality/Model management procedure, including functionality/model selection/activation/deactivation, and functionality/model switching/fallback/transfer/delivery/update

- Latency/interruption requirement for above procedures

The following LCM related requirements can be considered:

- Model/Functionality select/switch/activate/deactivate/fallback

- Model/Functionality monitoring

- On whether requirements for data collection (in particular for training) could/need be defined:

- Data collection requirements would only be defined if data collection procedure is defined in 3GPP specifications.

- On requirements for model transfer/update:

- Requirements would only be defined if model transfer/update would be defined in 3GPP specifications.

The legacy framework for RRC/MAC-CE/DCI based core requirements (e.g., define delay requirements based on multiple delay components) can be used as the baseline for LCM procedures if the LCM related requirements are agreed to be introduced. If new procedures which legacy framework is not applicable to are introduced, additional core requirement framework can be discussed.

LCM related tests should consider how the framework can address the possibility of updates/activation/deactivation /switching to the functionalities/models after the deployment of the devices in the field.

#### 7.3.2.3 Reference block diagrams for testing

Reference block diagrams provide test modules/functionalities of TE/DUT and testing framework for different use cases. Both reference block diagrams for 1-sided model and 2-sided model are studied.

##### 7.3.2.3.1 Reference block diagram for 1-sided model

Figure 7.3.2.3-1 provides the reference block diagram for 1-sided model. LCM in the figure includes functionality and/or model ID based LCM. The link between TE and DUT are physical and not logical. The logical link will depend on the functionality being tested. The scope of the figure includes both performance and potentially LCM tests. Offline training is assumed and some blocks may not be used in some of the tests. LCM may not be tested depending on the purpose of the test.

LCM

Verification

AI/ML functions

Test configuration/controller

Signal generator

inference

LCM

DUT

TE

Figure 7.3.2.3-1: Reference block diagram for 1-sided model

##### 7.3.2.3.2 Reference block diagram for 2-sided model

Figure 7.3.2.3-2 provides the reference block diagram for 2-sided model. LCM in the figure includes functionality and/or model ID based LCM. The link between TE and DUT are physical and not logical. The logical link will depend on the functionality being tested. The scope of the figure includes both performance and potentially LCM tests. Offline training is assumed and some blocks may not be used in some of the tests. LCM may not be tested depending on the purpose of the test.

LCM

Verification

AI/ML functions

Test configuration/controller

Signal generator

inference

LCM

DUT

TE

inference

Figure 7.3.2.3-2: Reference block diagram for 2-sided model

#### 7.3.2.3 Test encoder/decoder for 2-sided model

(NOTE: At the current stage the framework in this session applies to CSI compression case.)

In order to determine the test encoder/decoder, the following issues are considered:

- Common assumptions for proposals of the test decoder / encoder (and the paired encoder/ decoder) for tester

- The need for and potential definition and derivation procedure of intermediate KPI for decoder evaluation and selection

- Data collection/generation for decoder evaluation, and the common assumptions/environment needed for data collection/generation

- How to minimize the impact of possible variations/differences in the test decoder/ test encoder design/implementation on UE/ gNB performance verification

- The impact of test decoder/ encoder for testing complexity to UE/gNB performance verification, and the advantage/disadvantage analysis of high/low complexity decoders.

The test decoder/encoder design should take into account complexity limitations based on e.g., feasibility of TE implementation and complexity levels considered feasible by network vendors/UE vendors for decoder/encoder deployment.

The choice of test decoder/encoder should aim as much as possible to avoid limiting the implementation choices, including e.g. complexity, back-bone model etc, of UE/gNB encoders/decoders operating in the field (this principle may not be fully achievable in practice).

Specification on the test may include some high-level parameters for the test decoder/encoder (e.g. parameters related to processing complexity, model structure, etc).

Following the above principles, the considered options of test decoder are listed below

- Option 1: DUT provides the decoder

- Option 2: Infra vendor provides the decoder

- Option 3: Full decoder specification in standard

- Option 4: TE vendor provides the decoder

Option 3 target is that a single decoder defined in the specifications for at least a single test for any DUTs.

For option 4, the following aspects should be considered

- TE vendor should be able to develop the decoder based on the specifications

- Test repeatability should be ensured (variation among TE vendor implementations should be bound)

- Other vendors should also be able to develop such a decoder and which can deliver similar performance

- Interoperability should be ensured based on the parameters that need to be specified

- Parameters that need to be specified are FFS

- Candidate parameters/conditions that may be considered for defining test decoder include

- Training data set for TE decoder training

- Model structure (Activation function is included in the model structure)

- Performance parameters for the TE decoder (e.g. cosine similarity, loss function, etc)

- Maximum FLOPs allowed for the test decoder

- Maximum number/size of model parameters

- Compression ratio of decoder (output size/input size)

- Quantization level

- Other parameters are not precluded and to be further discussed.

- Note: Feasibility of definition of parameters needs further investigated.

Option 4 target is that a single decoder implemented by each TE vendor will be enough for at least a single test for any DUTs. TE vendor should be able to implement the test decoder for Option 4 without any involvement from another party. If this is found infeasible, another option in which TE vendors need to collaborate with DUT/infra vendors to implement the decoder could be considered.

Further clarifications and analysis of the four options of test decoder are included in Table 7.3.2.3-1. It is assumed that for Option 4 the TE vendors can implement the decoder just based on the specifications (no other party involved). The table would need to be revised if collaboration between TE vendor and DUT/infra vendor is needed.

Table 7.3.2.3-1: Comparison of the four options of test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendors (no need to be specified) | Up to decoder implementer (infra vendor) | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFSCould be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) |  |  |  |  |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TENeed to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long a the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| **Pros/Cons analysis** |
| Reflection on the real deployment (likelihood that test decoder would be used |  |  |  |  |
| TE requirements to deploy the decoder (e.g., training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TELower thank Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TELower thank Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TELower thank Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TEHigher than Option 3 in terms of that training at TE is required Note: How to ensure compatibility/ interoperability between TE and DUT needs further study |
| Specification effort (defining test decoder and requirements) | Low | Low | HighestRAN4 needs to standardize the entire decoder | HighRAN4 needs to study and may decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure (after specs are published) |  |  | No | No |
| Applicability to different scenarios/conditions/ configurations |  |  |  |  |
| Complexity of testing for the ecosystem | Testing the encoder at DUTHigher than Option 3/4 Need for interaction between TE vendor | Testing the encoder at DUTHigher than Option 3/4 Testing complexity higher also than Option 1 | Testing the encoder at DUTLow – no need for interaction between TE vendors and other parties | Testing the encoder at DUTLow – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than Option 3/4 FSS compared to Option 2 | Higher than Option 3/4 FSS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem |  |  |  |  |
| Friendly to STOA (state of the art) model test / Forward compatibility when new AI models are invented |  |  |  |  |
| Relationship with reference decoder/encoder (used by RAN4 to define the performance requirements) for defining the requirement |  |  |  |  |
| Whether model transfer/delivery is needed during the test procedure |  |  |  |  |

The feasibility of any of the testing options has not concluded and more study is required. Other testing options are not precluded and different options might be required after RAN4 performs additional studies.

#### 7.3.2.4 Data collection/generation for testing

Different generating methods of test dataset can be used for different tests. The following candidate methods are to be considered:

- Dataset based on TR 38.901, e.g. UMa channel, UMi channel, CDL channel, "legacy approach", etc.

- "Legacy approach" refers legacy test in which a channel model is used

- Field dataset (data collected directly from field measurements)

- TE generates dataset for test based on assumptions/parameters defined by RAN4 (e.g. by defining some rules/function to generate data)

- Other methods are not precluded

#### 7.3.2.5 Data collection/generation for training

Some conditions and/or accuracy requirements for the training dataset or training data generation could only be introduced if the training procedure is defined in 3GPP specifications.

#### 7.3.2.6 Generalization/scalability aspects

The necessity and feasibility of defining requirements or test to verify the generalization of AI/ML is studied.

The goals of generalization test are to verify whether the minimum level of performance of AI/ML functionality/model can be achieved/maintain under the identified scenarios and/or configurations, while the performance won’t be significantly degraded in other scenarios and/or configurations. The following aspects should be considered for generalization/scalability related testing:

- details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality

- what the minimum level performance for each identified scenario and/or configuration is

- what the significant degradation for other scenarios and/or configurations is

It should also be considered that generalization and/or scalability related requirements for different scenarios/ configurations can be implicitly handled in the test case definition.

As for the handling of generalization tests, the following option is considered as baseline:

Signalling based LCM procedures and performance monitoring are considered in dedicated test cases and are excluded in tests verifying generalization. RAN4 may define multiple tests with different conditions. In each of the test, TE configures the same specified UE configuration, and therefore the same specified UE configuration is tested under different conditions to verify its generalizability. (environment differs in each test but not changing dynamically during the test)

- Specified UE configuration includes functionality and/or model ID if defined

#### 7.3.2.7 AI/ML processing capability

The practical processing capability and implementation complexity for device under test should be assumed when specifying RAN4 requirements.

- The UE capability may be needed to handle different complexity for one side and two-side models.

- The complexity of UE should also be studied when making assumption on gNB side model, and vice versa.

### 7.3.3 CSI feedback enhancement

Both time domain CSI prediction and spatial-frequency domain CSI compression are considered.

PMI reporting framework (follow PMI vs. random PMI test, use of γ as criteria, etc.) is taken as starting point for CSI related tests. Other metrics/framework is not precluded.

For metrics for CSI requirements/tests, the following test metrics are identified:

- Option 1: Throughput/relative throughput

- Option 2: SGCS, NMSE

- Option 3: CSI prediction accuracy

Option 1 should be used as baseline. For option 3, further discuss is needed on the feasibility to define the CSI prediction accuracy in WI. For metrics for CSI monitoring, further discussion is needed in WI.

### 7.3.4 Beam management

Both spatial-domain DL beam prediction and temporal DL beam prediction are considered.

For metrics for beam management requirements/tests, the following test metrics are identified and could be considered

- Option 1: RSRP accuracy

- Option 2: Beam prediction accuracy

- Top-1 (%) : the percentage of "the Top-1 strongest beam is Top-1 predicted beam"

- Top-K/1 (%) : the percentage of "the Top-1 strongest beam is one of the Top-K predicted beams"

- Top-1/K (%) : the percentage of "the Top-1 predicted beam is one of the Top-K strongest beams"

- Option 3: The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB,

- Related measurement accuracy can be considered to determine x

- Option 4: combinations of above options

The overhead/latency reduction should be considered for the requirements as the side condition.

### 7.3.5 Positioning accuracy enhancements

Both direct AI/ML positioning and AI/ML assisted positioning are considered.

For metrics for positioning requirements/tests, the candidate options include

- Option 1: positioning accuracy: Ground truth vs. reported

- only option available for direct positioning

- Option 2: CIR/PDP, channel estimation accuracy

- Option 3: ToA, RSTD and RSRP, and RSRPP

- Option 4: others (e.g., intermediate KPIs, LoS/NLoS)/combinations of the above

The feasibility and testability of different options should be further justified in WI.

# 8 Conclusions

The following aspects have been studied for the general framework of AI/ML over air interface for one-sided models and two-sided models:

- Various Network-UE Collaboration Levels

- Functionality-based LCM and model-ID-based LCM

- Functionality/model selection, activation, deactivation, switching, fallback

- Functionality identification and model identification

- Data collection

- Performance monitoring

- Various model identification Types and their use cases

- Reporting of applicable functionalities/models

- Method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE

- Model delivery/transfer and analysis of various model delivery/transfer Cases

The above studied aspects for General Framework can be considered for developing/specifying AI/ML use cases and common framework (if needed for some aspects) across AI/ML use cases.

***CSI feedback enhancement:***

**CSI compression sub use case:**

The performance benefit and potential specification impact were studied for AI/ML based CSI compression sub use case.

Evaluation has been performed to assess AI/ML based CSI compression from various aspects, including performance gain over non-AI/ML benchmark, model input/output type, CSI feedback quantization methods, ground-truth CSI format, monitoring, generalization, training collaboration types, etc. Some aspects were studied but not fully investigated, including the options of CQI/RI calculation, the options of rank>1 solution.

Performance gain over baseline and computational complexity in FLOPs are summarized in clause 6.2.2.8.

Potential specification impact on NW side/UE side data collection, dataset delivery, quantization alignment between CSI generation part at the UE and CSI reconstruction part at the NW, CSI report configuration, CSI report format, pairing information/procedure and monitoring approach were investigated but not all aspects were identified.

The pros and cons are analysed for each training collaboration types, and each training collaboration type has its own benefits and limitations in different aspects. The study has investigated the feasibility of the studied training collaboration types and necessity of corresponding potential RAN1 specification impact. However, not all aspects have been concluded.

Both NW side and UE side performance monitoring were studied, some but not all aspects were concluded.

From RAN1 perspective, there is no consensus on the recommendation of CSI compression for normative work.

At least the following aspects are the reasons for the lack of RAN1 consensus on the recommendation of CSI compression for normative work:

- Trade-off between performance and complexity/overhead.

- Issues related to inter-vendor training collaboration.

Other aspects that require further study/conclusion are captured in the summary above.

**CSI prediction sub use case:**

The performance and potential specification impact were studied for AI/ML based UE side CSI prediction sub use case.

Evaluations have been performed to assess AI/ML based CSI prediction from various aspects, including performance compared to baseline, model input/output type, generalization over UE speed, etc. Some aspects are studied but lack observations, including scalability over various configurations and generalization over other scenarios and approach of fine tuning. Performance monitoring accuracy has not been evaluated.

Performance compared with baseline is summarized in clause 6.2.2.8.

Potential specification impact on data collection and performance monitoring are discussed in clause 7.2.2. Limited specification aspects were considered.

From RAN1 perspective, there is no consensus on the recommendation of CSI prediction for normative work.

The reason for the lack of RAN1 consensus on the recommendation of CSI prediction for normative work:

- Lack of results on the performance gain over non-AI/ML based approach and associated complexity.

Other aspects that require further study/conclusion are captured in the summary above.

***Beam management:***

This study focuses on evaluation of potential benefits of AI/ML-based beam management and analysis of potential enhancements to enable AI/ML for beam management.

During the study, BM-Case1 (Spatial-domain downlink beam prediction) and BM-Case2 (Temporal-domain downlink beam prediction), as described in clause 5.2, are selected as the representative sub use cases.

Evaluation scenarios and KPIs are described in clause 6.3.1, and the detailed evaluation results from different sources and the key observations are captured in clause 6.3.2. Evaluation results have shown that it is beneficial to enable AI/ML for beam management in the considered evaluation scenarios.

The necessity, feasibility, benefit and potential specification impacts of potential enhancements to enable AI/ML for beam management were studied from different aspects, and the outputs are captured in clause 7.

For AI-based beam management, from RAN1 perspective, at least the following are recommended for normative work:

- Both BM-Case1 and BM-Case2:

- BM-Case1: Spatial-domain DL Tx beam prediction for Set A of beams based on measurement results of Set B of beams

- BM-Case2: Temporal DL Tx beam prediction for Set A of beams based on the historic measurement results of Set B of beams

- DL Tx beam prediction for both UE-sided model and NW-sided model

- Necessary signalling/mechanism(s) to facilitate data collection, model inference, and performance monitoring for both UE-sided model and NW-sided model

- Signalling/mechanism(s) to facilitate necessary LCM operations via 3GPP signalling for UE-sided model

***Positioning accuracy enhancements:***

This study focused on the analysis of potential enhancements necessary to enable AI/ML for positioning accuracy enhancements with NR RAT-dependent positioning methods.

Evaluation scenarios and KPIs were identified for system level analysis of AI/ML enabled RAT-dependent positioning techniques as described in clause 6.4.

Direct AI/ML positioning and AI/ML assisted positioning were identified and selected as the representative sub-use cases. Evaluation results have shown that in considered evaluation scenarios (i.e., InF-DH, and other InF scenarios), both direct AI/ML positioning and AI/ML assisted can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods. Various aspects of AI/ML for positioning accuracy enhancement were investigated and evaluated as described in clause 6.4 that provides summary of evaluation results from different sources.

Based on the conducted analysis, it is recommended to proceed with normative work for AI/ML based positioning.

The necessity, feasibility and potential enhancements to facilitate the support of AI/ML for positioning accuracy enhancements with NR RAT-dependent positioning methods were studied and the outcome are outlined in clause 7.

It is recommended to specify necessary measurement, signalling and procedure to facilitate training, inference, monitoring and/or other LCM operations for both direct AI/ML positioning and AI/ML assisted positioning, specifically:

- specify necessary signalling of data collection; investigate the necessity of other information for supporting data collection, and if needed, specify during normative work

- investigate on the necessity and signalling details of measurement enhancements, and if needed, specify during normative work

- investigate on the necessity and signalling details of monitoring method(s), and if needed, specify during normative work

A variety of enhancements for measurements (e.g., based on extensions to current positioning measurements or with new measurements) were also identified as potentially beneficial (e.g., trade-off positioning accuracy requirement and signalling overhead) and are recommended to be investigated further and if needed, specified during normative work.

Annex A: Change history

|  |
| --- |
| **Change history** |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2022-05 | RAN1#109e |  |  |  |  | TR skeleton | 0.0.0 |
| 2023-05 | RAN1#113 | R1-2306235 |  |  |  | RAN1 agreement up to and including RAN1#112bis-e  | 0.1.0 |
| 2023-08 | RAN1#114 | R1-2308681 |  |  |  | RAN1 agreements from RAN1#113 and RAN1#114 | 0.2.0 |
| 2023-09 | RAN#101 | RP-231766 |  |  |  | TR presented for information at RAN#101 [same as R1-2308681] | 1.0.0 |
| 2023-10 | RAN1#114b | R1-2310163 |  |  |  | Added clause numbering for CSI and BM performance results.  | 1.1.0 |
| 2023-11 | RAN1#115 | R1-2312055 |  |  |  | RAN1 agreements from RAN1#114b | 1.2.0 |
| 2023-11 | RAN1#115 | R1-2312764 |  |  |  | Agreements from RAN1#115, RAN2#124 and RAN4#109 | 1.3.0 |
| 2023-12 | RAN#102 | RP-233133 |  |  |  | TR presented for approval at RAN#102 [same as R1-2312764] | 2.0.0 |
| 2023-12 | RAN#102 | RP-233946 |  |  |  | Editorial corrections for RAN approval, removal of square brackets in Scope clause.  | 2.0.1 |
| 2023-12 | RAN#102 |  |  |  |  | Rel-18 TR under change control further to RAN approval | 18.0.0 |