**3GPP TSG-RAN WG2 Meeting #123 *R2-23XXXXX***

**Toulouse, France, August 21 – 25, 2023**

|  |
| --- |
| *CR-Form-v12.2* |
| **CHANGE REQUEST** |
|  |
|  | **38.843** | **CR** | **-** | **rev** | **-** | **Current version:** | **0.1.0** |  |
|  |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME | **X** | Radio Access Network | **x** | Core Network |  |

|  |
| --- |
|  |
| ***Title:***  | R2 input to TR 38.343  |
|  |  |
| ***Source to WG:*** | Ericsson |
| ***Source to TSG:*** | R2 |
|  |  |
| ***Work item code:*** | FS\_NR\_AIML\_air |  | ***Date:*** | 2023-08-11 |
|  |  |  |  |  |
| ***Category:*** | **B** |  | ***Release:*** | Rel-18 |
|  | *Use one of the following categories:****F*** *(correction)****A*** *(mirror corresponding to a change in an earlier release)****B*** *(addition of feature),* ***C*** *(functional modification of feature)****D*** *(editorial modification)*Detailed explanations of the above categories canbe found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | *Use one of the following releases:Rel-8 (Release 8)Rel-9 (Release 9)Rel-10 (Release 10)Rel-11 (Release 11)…Rel-16 (Release 16)Rel-17 (Release 17)Rel-18 (Release 18)Rel-19 (Release 19)* |
|  |  |
| ***Reason for change:*** | Introduce R2 agreements to the Technical Report |
|  |  |
| ***Summary of change:*** | * §4.2: Adding Editor’s Notes / R2-centric comments
* §4.4: Introducing functional framework details
* §7.3: Related Editor’s Note
* §7.3.1: Subdividing the “Common framework” clasue as follows…
* §7.3.1.1: Adding “Model and Functionality Identification” subclause
* §7.3.1.2: Adding “Data collection” subclause
* §7.3.1.3: Adding “Model Transfer/Delivery” subclause
* §7.3.1.4: Placeholder for “UE Capability Reporting” subclause
* §7.3.1.5: Placeholder for “Applicability Reporting” subclause
* §7.3.2: Adding input to “CSI feedback enhacement” clause
* §7.3.3: Adding input to “Beam management” clause
* §7.3.4: Adding input to “Positioning accuracy enhancement” clause
 |
|  |  |
| ***Consequences if not approved:*** | No R2 protocol related aspects included in the TR. |
|  |  |
| ***Clauses affected:*** | 4.2, 4.4, 7.3, 7.3.1, 7.3.1.1, 7.3.1.2, 7.3.1.3, 7.3.1.4, 7.3.2, 7.3.3, 7.3.4 |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  | **X** |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  | **X** |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  | **X** |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
| ***This CR's revision history:*** |  |

|  |
| --- |
| 3GPP TR 38.843 V0.1.0 (2023-05) |
| Technical Report |
| 3rd Generation Partnership Project;Technical Specification Group Radio Access Networks;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.© 2022, 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 4

Introduction 5

1 Scope 6

2 References 6

3 Definitions of terms, symbols and abbreviations 7

3.1 Terms 7

3.2 Symbols 8

3.3 Abbreviations 9

4 General AI/ML Framework 9

4.1 Description of AI/ML stages 9

4.2 AI/ML model Life Cycle Management 9

4.3 Collaboration levels 11

4.4 Functional Framework Details 11

5 Use cases 11

5.1 CSI feedback enhancement 12

5.2 Beam Management 12

5.3 Positioning accuracy enhancements 13

6 Evaluations 14

6.1 Common evaluation methodology and KPIs 14

6.2 CSI feedback enhancement 15

6.2.1 Evaluation assumptions, methodology and KPIs 15

6.2.2 Performance results 24

6.3 Beam Management 35

6.3.1 Evaluation assumptions, methodology and KPIs 35

6.3.2 Performance results 42

6.4 Positioning accuracy enhancements 44

6.4.1 Evaluation assumptions, methodology and KPIs 44

6.4.2 Performance results 50

7 Potential Specification Impact Assessment 52

7.1 General observations 52

7.2 Physical layer aspects 53

7.2.1 Common framework 53

7.2.2 CSI feedback enhancement 53

7.2.3 Beam management 55

7.2.4 Positioning accuracy enhancements 56

7.3 Protocol aspects 59

7.3.1 Common framework 59

7.3.2 CSI feedback enhancement 59

7.3.3 Beam management 59

7.3.4 Positioning accuracy enhancements 59

7.4 Interoperability and testability aspects 59

7.4.1 Common framework 59

7.4.2 CSI feedback enhancement 59

7.4.3 Beam management 59

7.4.4 Positioning accuracy enhancements 59

8 Conclusions 59

Annex <X> : Change history 60

For definitive guidance on drafting 3GPP TSs and TRs, see [3GPP TS 21.801](http://www.3gpp.org/DynaReport/21801.htm) supplemented by the 3GPP web page <http://www.3gpp.org/specifications-groups/delegates-corner/writing-a-new-spec>.

Ensure all blue guidance text is removed before submitting the TS/TR to the TSG for approval.

# Foreword

This clause is mandatory; do not alter the text in any way other than to choose between "Specification" and "Report".

This Technical Specification|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 drafting the TS/TR, pay particular attention to the use of modal auxiliary verbs! TRs shall not contain any normative provisions.

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.

# Introduction

This clause is optional. If it exists, it shall be the second unnumbered clause.

# 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".

…

[x] <doctype> <#>[ ([up to and including]{yyyy[-mm]|V<a[.b[.c]]>}[onwards])]: "<Title>".

It is preferred that the reference to 21.905 be the first in the list.

# 3 Definitions of terms, symbols and abbreviations

This clause and its three subclauses are mandatory. The contents shall be shown as "void" if the TS/TR does not define any terms, symbols, or abbreviations.

## 3.1 Terms

For the purposes of the present document, the terms given in 3GPP 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 3GPP TR 21.905 [1].

**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.

**Model activation:** enable an AI/ML model for a specific function.

**Model deactivation:** disable an AI/ML model for a specific function.

**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 function.

**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. (This note could be removed when we define the term fine-tuning.)

**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*.

**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 format (EW)

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP 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 3GPP TR 21.905 [1].

AI Artificial Intelligence

BM Beam Management

CIR Channel Impulse Response

CNN Convolutional Neural Network

CSI Channel State Information

DL Downlink

EVM Evaluation Methodology

FLOPS Floating Point per Second

GCS Generalized Cosine Similarity

KPI Key Performance Indicator

LCM Life Cycle Management

LLS Link Level Simulations

ML Machine Learning

NMSE Normalized Mean Square Error

PDP Power Delay Profile

RNN Recurrent Neural Network

SGCS Squared Generalized Cosine Similarity

SLS System Level Simulations

UPT User Perceived Throughput

# 4 General AI/ML Framework

The purpose of this section 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 section, 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.]

*Editor’s notes: This section should cover the introduction model training, model inference, model monitoring. FL to have a* ***figure*** *for description. Each box has a one-liner description with details elaborated in section 4.4.*

## 4.2 AI/ML model Life Cycle Management

*Editor’s note: To discuss the changes needed in this section to reflect the option of AI/ML functionality-based LCM.*

*Editor’s note: This section should be updated to align with what clause 4.4. describes.*

In this section, the lifecycle management of AI/ML model is characterized, e.g., model training, model deployment, model inference, model monitoring, model updating.

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

* Data collection
	+ Note: This also includes associated assistance information, if applicable.
* Model training
* Functionality/model identification
* Model 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
	+ Note: Terminology is to be defined. This includes model finetuning, retraining, and re-development via online/offline training.
* UE capability

Notes: Some aspects in the list may not have specification impact. Aspects with square brackets are tentative and pending terminology definition. More aspects may be added as study progresses.

The LCM procedure is studied on the basis that an AI/ML model has a *model ID* with associated information and/or *model functionality* at least for some AI/ML operations.

=====

*Editor’s note: consider breaking paragraphs below into new subsection under 4.2 (possibly above too).*

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

* For *AI/ML functionality identification*
	+ Reuse legacy 3GPP framework of Features as a starting point for discussion.
	+ 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 will be studied. 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.

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.

 *Editor’s note: RAN2 should address and study impact on RRC protocol, including UE capability reporting and other related signalling.*

After *functionality identification*, necessity, mechanisms, for UE to report updates on applicable functionality(es) among [configured/identified] functionality(es), where the applicable functionalities may be a subset of all [configured/identified] functionalities are studied.

*Editor’s note: RAN2 to discuss in this section technical inputs related to reporting updates to the applicability of functionalities.*

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. Model ID can also be used for model identification between the Network and the UE, and model transfer / delivery (if supported).

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.

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.

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

*Editor’s note: RAN2 to discuss in this section technical inputs related to reporting updates to the applicability of models.*

*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.

*Editor’s note: Details on data collection should later be aligned according to Clause 4.4.*

## 4.3 Collaboration levels

In this section, various levels of collaboration between UE and gNB 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/gNB 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.

The following Cases further detail the 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:

|  |  |  |  |
| --- | --- | --- | --- |
| **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 | 3GPP Network | NW-side |
| **z5** | model transfer in open format of an unknown model structure at UE | 3GPP Network | NW-side |

## 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 to address both model-based as functionality-based LCM introduced in clause 4.2. For the functions and data/information flows shown in the Figure 4.4-1, whether there is any standardization impact and what is the standardization impact are discussed in clause 7.



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

*Editor’s note: The need/purpose of the different data/information flows (i.e., arrows) should be further clarified.*

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

* 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.
* The Model Training function performs the AI/ML model training, validation, and testing which may generate model performance metrics 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: Used to send trained, validated, and tested AI/ML models to the Model Storage function (if any), or to send an updated version of a model to the Model Storage function (if any).
* Model Management is a function that oversees the operation and monitoring 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 for the Inference function to fine-tune its operation. 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.
	+ Monitoring output: Monitoring output used for the (re)training purposes.
* Inference is a function that provides outputs from the process of applying AI/ML models or AI/ML functionalities to new data (i.e., Inference Data) to perform predictions or decisions. 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 models that can be used to perform the inference process.

	+ Note: The Model Storage function, if any, is primarily intended as a reference point 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.
	+ 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 sided model

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 network-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 [, or parallel training] at UE and NW
* 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.

[Pros/cons of different offline training collaboration types are analyzed with respect to the following aspects:

* Whether model can be kept proprietary
* Requirements on privacy-sensitive dataset sharing
* Flexibility to support cell/site/scenario/configuration specific model
* gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware
* Model update flexibility after deployment
* feasibility of allowing UE side and NW side to develop/update models separately
* Model performance based on evaluation in 9.2.2.1
* Whether gNB can maintain/store a single/unified model
* Whether UE device can maintain/store a single/unified model
* Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use
* Whether training data distribution can be matched to the device that will use the model for inference
* Whether device capability can be considered for model development
* Other aspects are not precluded
* Note: training data collection and dataset/model delivery will be discussed separately]

## 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 and Set B is for DL beam measurement. The beam patterns of Set A and Set B can be clarified by companies.
	+ AI/ML model input: 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: 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.
	+ [AI/ML model output]: F predictions for F future time instances, 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 for [AI/ML model output] are defined:

* 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. Values 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.

## 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.

# 6 Evaluations

In this section, 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 [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
	+ 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

## 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.

For the evaluation of the AI/ML based CSI feedback enhancement, for ‘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.

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 model size, number of AI/ML parameters
	+ 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.
* AI/ML memory storage in terms of AI/ML model size and number of AI/ML parameters is adopted as part of the ‘Evaluation Metric’, and reported by companies who may select either or both.
* 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: monitoring mechanism 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: 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.
			* Option 2: 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* can be the same or different from 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*:**

In order to study the verification of generalization, the following aspects are encouraged to be reported:

* The configuration(s)/scenario(s) for training dataset, including potentially the mixed training dataset from multiple configurations/scenarios
* The configuration(s)/scenario(s) for testing/inference
* The detailed list of configuration(s) and/or scenario(s)

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#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.
	+ 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, FFS candidate payload number
* 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.
* Companies to report the selected configurations for generalization verification
* Companies are encouraged to report 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.

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

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.

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
* 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
* 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
	+ 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.
* 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
	+ 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.
* 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)

***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.FFS ideal 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 rank

Additional 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 taken 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: Companies need to report which option is used between:* 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 reported

Note: 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, FFS ideal channel estimation |
| Rank per UE | Rank 1-4. Companies are encouraged to report the Rank number, and whether/how rank adaptation is applied |

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

For the evaluation of the AI/ML based **CSI compression** sub use case, companies are encouraged to report details of their models, including:

* 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:
	+ Raw channel matrix (in frequency or delay domain), e.g., channel matrix with dimensions of Tx, Rx, and frequency unit
	+ Precoding matrix (as a group of eigenvectors or an eTypeII-like reporting)
* Data pre-processing/post-processing
* Loss function
* Specific quantization/dequantization method, e.g., vector quantization, scalar quantization, etc, considering the following aspects:
	+ Quantization non-aware training, where the float-format variables are directly passed from CSI generation part to CSI reconstruction part during the training
		- 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
	+ Quantization-aware training, where quantization/dequantization is involved in the training process
		- 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.
		- 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
* Considering performance impact of ground truth quantization in the CSI compression
	+ Studying 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
* 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.
		- Note: input/output type is Precoding matrix
		- 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.
		- Note: input/output type is Precoding matrix
		- 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

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

For the evaluation of the AI/ML based **CSI prediction** sub use case, companies are encouraged to report details of their models, including:

* 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 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.

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

### 6.2.2 Performance results

Tables 6.2.2-1 through 6.2.2-4 present the performance results for the evaluation results of AI/ML-based CSI compression without and with generalization/scalability verification for different training assumptions, namely, 1-on-1 joint training, multi-vendor joint training and separate training.

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

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
	+ Other high resolution CSI quantization methods can be additionally submitted for comparison, e.g., R16 Type 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

 Table 6.2.2-1: Evaluation results for CSI compression 1-on-1 joint training without model generalization/scalability, [traffic type], [Max rank value], [RU]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| CSI generation part | AL/ML model backbone |  |  |
| Pre-processing |  |  |
| Post-processing |  |  |
| FLOPs/M |  |  |
| Number of parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| CSI reconstruction part | AL/ML model backbone |  |  |
| [Pre-processing] |  |  |
| [Post-processing] |  |  |
| FLOPs/M |  |  |
| Number of parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| Common description | Input type |  |  |
| Output type |  |  |
| Quantization /dequantization method |  |  |
| Rank/layer adaptation settings for rank>1 |  |  |
| Dataset description | Train/k |  |  |
| Test/k |  |  |
| Ground-truth CSI quantization method (incl. scalar/codebook based quantization, and the parameters) |  |  |
| Overhead reduction compared to Float32 if high resolution quantization of ground-truth CSI is applied |  |  |
| [Other assumptions/settings agreed to be reported] |  |  |
| Benchmark |  |  |
| Benchmark assumptions, e.g., CSI overhead calculation method (Optional) |  |  |
| SGCS of benchmark, [layer 1] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| SGCS of benchmark, [layer 2] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| Gain for SGCS, [layer 1] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| Gain for SGCS, [layer 2] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| …(other layers) |  |  |  |
| NMSE of benchmark, [layer 1] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| NMSE of benchmark, [layer 2] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| Gain for NMSE, [layer 1] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| Gain for NMSE, [layer 2] | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| …(other layers) |  |  |  |
| Other intermediate KPI (description/value) (optional) |  |  |
| Gain for other intermediate KPI (description/value) (optional) |  |  |
| Gain for Mean UPT (for a specific CSI feedback overhead) | CSI feedback payload A |  |  |
| CSI feedback payload B  |  |  |
| CSI feedback payload C |  |  |
| Gain for 5% UPT | CSI feedback payload A |  |  |
| CSI feedback payload B |  |  |
| CSI feedback payload C |  |  |
| Gain for upper bound without CSI compression over Benchmark – Mean UPT (Optional) | CSI feedback payload A |  |  |
| CSI feedback payload B |  |  |
| CSI feedback payload C |  |  |
| Gain for upper bound without CSI compression over Benchmark – 5% UPT (Optional) | CSI feedback payload A |  |  |
| CSI feedback payload B |  |  |
| CSI feedback payload C |  |  |
| [CSI feedback reduction (%)] |  |  |
| … |  |  |  |
| FFS others |  |  |  |

where, for Max rank = 1 or 2: X ≤ 80 bits; Y = 100 bits – 140 bits; Z ≥ 230 bits and for Max rank = 3 or 4, X ≤ $\left⌈160/υ\right⌉$ bits; Y = $\left⌈200/υ\right⌉$ bits – $\left⌈280/υ\right⌉ $bits; Z ≥ $\left⌈460/υ\right⌉$ bits.

where, CSI feedback payload A ≤ β∙80 bits; B = β∙(100 bits – 140 bits); C ≥ β∙230 bits. Note: β=1 for Max rank = 1 and β = 1.5 for Max rank = 2, 3 or 4.

The CSI feedback reduction is provided for three CSI feedback overhead ranges (RU ≤ 39%, 40% ≤ RU ≤ 69%, RU ≥ 70%) , 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. Note: the CSI feedback overhead reduction and gain for mean/5%tile UPT are determined at the same payload size for benchmark scheme.

Notes: “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/dequantizaion (SQ/VQ), etc. “Input type” means the input of the CSI generation part. “output type” means the output of the CSI reconstruction part.

Table 6.2.2-2: Evaluation results for CSI compression with model generalization/scalability, [Max rank value], [Scenario/configuration]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| CSI generation part | AL/ML model backbone |  |  |
| Pre-processing |  |  |
| Post-processing |  |  |
| FLOPs/M |  |  |
| Number of parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| CSI reconstruction part | AL/ML model backbone |  |  |
| [Pre-processing] |  |  |
| [Post-processing] |  |  |
| FLOPs/M |  |  |
| Number of parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| Common description | Input type |  |  |
| Output type |  |  |
| Quantization /dequantization method |  |  |
| Generalization/Scalability method description if applicable, e.g., truncation, adaptation layer, etc. |  |  |
| Input/output scalability dimension if applicable, e.g., N>=1 NW part model(s) to M>=1 UE part model(s) |  |  |
| Dataset description | Ground-truth CSI quantization method |  |  |
| [Other assumptions/settings agreed to be reported] |  |  |
| Generalization Case 1 | Train (setting#A, size/k) |  |  |
| Test (setting#B, size/k) |  |  |
| SGCS, layer 1 | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| SGCS, layer 2 | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| NMSE, layer 1 | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| NMSE, layer 2 | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| …(other settings for Case 1) |  |  |  |
| … |  |  |  |
| Generalization Case 2 | Train (setting#A, size/k) |  |  |
| Test (setting#B, size/k) |  |  |
| …(results for Case 2) |  |  |  |
| …(other settings for Case 2) |  |  |  |
| Generalization Case 3 | Train (setting#A, size/k) |  |  |
| Test (setting#B, size/k) |  |  |
| …(results for Case 3) |  |  |  |
| Fine-tuning case (optional) |  |  |  |
| …(results for Fine-tuning) |  |  |  |
| …(other settings for Fine-tuning) |  |  |  |
| FFS others |  |  |  |

Notes: “Quantization/dequantization method” includes the description of training awareness (Case 1/2-1/2-2), type of quantization/dequantizaion (SQ/VQ), etc. “Input type” means the input of the CSI generation part. “output type” means the output of the CSI reconstruction part.

The intermediate KPI results are in the form of absolute value and the gain over a given benchmark, e.g., in terms of “absolute value (gain over benchmark)”. SGCS is to be expressed in linear domain, while NMSE in dB domain.

Table 6.2.2-3: Evaluation results for CSI compression of multi-vendor joint training without model generalization/scalability, [traffic type], [Max rank value], [RU]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| Common description | Input type |  |  |
| Output type |  |  |
| [Training method] |  |  |
| Quantization /dequantization method |  |  |
| Dataset description | Train/k |  |  |
| Test/k |  |  |
| Ground-truth CSI quantization method (incl. scalar/codebook based quantization, and the parameters) |  |  |
| Case 1 (baseline): NW#1-UE#1 | UE part AI/ML model backbone/structure |  |  |
| Network part AI/ML model backbone/structure |  |  |
| ...(other NW-UE combinations for Case 1) |  |  |  |
| Case 2 (1 NW part to M>1 UE parts) | NW part model backbone/structure |  |  |
| UE#1 part model backbone/structure |  |  |
| UE#1 part training dataset description and size |  |  |
| … |  |  |
| UE#M part model backbone/structure |  |  |
| UE#M part training dataset description and size |  |  |
| Case 3 (N>1 NW parts to 1 UE part) | UE part model backbone/structure |  |  |
| NW#1 part model backbone/structure |  |  |
| NW#1 part training dataset description and size |  |  |
| … |  |  |
| NW#N part model backbone/structure |  |  |
| NW#N part training dataset description and size |  |  |
| Intermediate KPI type (SGCS/NMSE) |  |  |
| FFS other cases |  |  |  |
| Case 1: NW#1-UE#1: Intermediate KPI | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| …(results for other NW-UE combinations for Case 1) |  |  |  |
| Case 2: Intermediate KPI | CSI feedback payload X,  |  |  |
| NW-UE#1 |  |  |
| … |  |  |
| CSI feedback payload X,  |  |  |
| NW-UE#M |  |  |
| Case 3: Intermediate KPI | CSI feedback payload X,  |  |  |
| NW#1-UE |  |  |
| … |  |  |
| CSI feedback payload X,  |  |  |
| NW#N-UE |  |  |
| FFS other cases |  |  |  |
| FFS others |  |  |  |

Notes: “Quantization/dequantization method” includes the description of training awareness (Case 1/2-1/2-2), type of quantization/dequantizaion (SQ/VQ), etc. “Input type” means the input of the CSI generation part. “output type” means the output of the CSI reconstruction par

The intermediate KPI results are in the form of absolute value and the gain over a given benchmark, e.g., in terms of “absolute value (gain over benchmark)”. SGCS is to be expressed in linear domain, while NMSE in dB domain.

Table 6.2.2-4: Evaluation results for CSI compression of separate training without model generalization/scalability, [Max rank value]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| Common description | Input type |  |  |
| Output type |  |  |
| Quantization /dequantization method |  |  |
| Shared output of CSI generation part/input of reconstruction part is before or after quantization |  |  |
| Dataset description | Test/k |  |  |
| Ground-truth CSI quantization method  |  |  |
| [Benchmark: NW#1-UE#1 joint training] | UE part AI/ML model backbone/structure |  |  |
| Network part AI/ML model backbone/structure |  |  |
| Training dataset size |  |  |
| ...(other NW-UE combinations for benchmark) |  |  |  |
| Case 1-NW first training | NW part model backbone/structure |  |  |
| UE#1 part model backbone/structure |  |  |
| UE#1 part training dataset description and size |  |  |
| … |  |  |
| UE#M part model backbone/structure |  |  |
| UE#M part training dataset description and size |  |  |
| [air-interface overhead of information (e.g., dataset) sharing] |  |  |
| Case 1-UE first training | NW#1 part model backbone/structure |  |  |
| NW#1 part training dataset description and size |  |  |
| … |  |  |
| NW#N part model backbone/structure |  |  |
| NW#N part training dataset description and size |  |  |
| UE part model backbone/structure |  |  |
| [air-interface overhead of information (e.g., dataset) sharing] |  |  |
| Case 2-UE first training | UE#1 part model backbone/structure |  |  |
| … |  |  |
| UE#M part model backbone/structure |  |  |
| UE part AI/ML model backbone/structure |  |  |
| NW part training dataset description and size (e.g., description/size of dataset from M UEs and how to merge) |  |  |
| Case 3-NW first training | NW#1 part model backbone/structure |  |  |
| … |  |  |
| NW#N part model backbone/structure |  |  |
| UE part model backbone/structure |  |  |
| UE part training dataset description and size (e.g., description/size of dataset from N NWs and how to merge) |  |  |
| Intermediate KPI type (SGCS/NMSE) |  |  |
| FFS other cases |  |  |  |
| NW#1-UE#1 joint training: Intermediate KPI | CSI feedback payload X |  |  |
| CSI feedback payload Y |  |  |
| CSI feedback payload Z |  |  |
| …(results for other 1-on-1 NW-UE joint training combinations) |  |  |  |
| Case 1-NW first training: Intermediate KPI | CSI feedback payload X, NW-UE#1 |  |  |
| … |  |  |
| CSI feedback payload X, NW-UE#M |  |  |
| CSI feedback payload Y … |  |  |
| CSI feedback payload Z … |  |  |
| Case 1-UE first training: Intermediate KPI | CSI feedback payload X, NW#1-UE |  |  |
| … |  |  |
| CSI feedback payload X, NW#N-UE |  |  |
| CSI feedback payload Y … |  |  |
| CSI feedback payload Z … |  |  |
| Case 2-NW first training: Intermediate KPI | CSI feedback payload X, NW#1-UE |  |  |
| … |  |  |
| CSI feedback payload X, NW#N-UE |  |  |
| CSI feedback payload Y … |  |  |
| CSI feedback payload Z … |  |  |
| Case 3-NW first training: Intermediate KPI | CSI feedback payload X, NW-UE#1 |  |  |
| … |  |  |
| CSI feedback payload X, NW-UE#M |  |  |
| CSI feedback payload Y …CSI feedback payload Z … |  |  |
|  |  |  |
| FFS other cases |  |  |  |
| FFS others |  |  |  |

Notes: “Quantization/dequantization method” includes the description of training awareness (Case 1/2-1/2-2), type of quantization/dequantizaion (SQ/VQ), etc. “Input type” means the input of the CSI generation part. “output type” means the output of the CSI reconstruction part.

The intermediate KPI results are in the form of absolute value and the gain over a given benchmark, e.g., in terms of “absolute value (gain over benchmark)”. SGCS is to be expressed in linear domain, while NMSE in dB domain.

Table 6.2.2-5 presents the performance results for the evaluation results of AI/ML-based CSI prediction without generalization/scalability verification.

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: 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: 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.

 Table 6.2.2-5: Evaluation results for CSI prediction without model generalization/scalability, [traffic type], [Max rank value], [RU]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| AI/ML model description | AL/ML model backbone |  |  |
| [Pre-processing] |  |  |
| [Post-processing] |  |  |
| FLOPs/M |  |  |
| Parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| Input type |  |  |
| Output type |  |  |
| Assumptions | UE speed |  |  |
| CSI feedback periodicity |  |  |
| Observation window (number/distance) |  |  |
| Prediction window (number/distance [between prediction instances/distance from the last observation instance to the 1st prediction instance]) |  |  |
| Whether/how to adopt spatial consistency |  |  |
| Codebook type for CSI report |  |  |
| Dataset size | Train/k |  |  |
| Test/k |  |  |
| Benchmark 1 |  |  |
| Intermediate KPI #1 of Benchmark 1 |  |  |  |
| Gain for intermediate KPI#1 over Benchmark 1 |  |  |  |
| Intermediate KPI #2 of Benchmark 1 |  |  |  |
| Gain for intermediate KPI#2 over Benchmark 1 |  |  |  |
| Gain for eventual KPI (Benchmark 1) | Mean UPT |  |  |
| 5% UPT |  |  |
| Benchmark 2 |  |  |
| Intermediate KPI #1 of Benchmark 2 |  |  |  |
| Gain for intermediate KPI#1 over Benchmark 2 |  |  |  |
| Intermediate KPI #2 of Benchmark 2 |  |  |  |
| Gain for intermediate KPI#2 over Benchmark 2 |  |  |  |
| Gain for eventual KPI (Benchmark 2) | Mean UPT |  |  |
| 5% UPT |  |  |
| FFS others |  |  |  |

Table 6.2.2-6 presents the performance results for the evaluation results of AI/ML-based CSI prediction with model generalization/scalability verification.

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.

Table 6.2.2-6: Evaluation results for CSI prediction with model generalization/scalability [Max rank value]

|  |  |  |
| --- | --- | --- |
|  | Source 1 | … |
| AI/ML model description | AL/ML model description (e.g., backbone, structure) |  |  |
| [Pre-processing] |  |  |
| [Post-processing] |  |  |
| FLOPs/M |  |  |
| Parameters/M |  |  |
| [Storage /Mbytes] |  |  |
| Input type |  |  |
| Output type |  |  |
| Assumptions | CSI feedback periodicity |  |  |
| Observation window (number/distance) |  |  |
| Prediction window (number/distance between prediction instances/distance from the last observation instance to the 1st prediction instanc) |  |  |
| Whether/how to adopt spatial consistency |  |  |
| Generalization Case 1 | Train (setting#A, size/k) |  |  |
| Test (setting#A, size/k) |  |  |
| SGCS (1,…N, N is number of prediction instances) |  |  |
| NMSE (1,…N, N is number of prediction instances) |  |  |
| …(other settings and results for Case 1) |  |  |  |
| Generalization Case 2 | Train (setting#A, size/k) |  |  |
| Test (setting#A, size/k) |  |  |
| SGCS (1,…N, N is number of prediction instances) |  |  |
| NMSE (1,…N, N is number of prediction instances) |  |  |
| …(other settings and results for Case 2) |  |  |  |
| Generalization Case 3 | Train (setting#A, size/k) |  |  |
| Test (setting#A, size/k) |  |  |
| SGCS (1,…N, N is number of prediction instances) |  |  |
| NMSE (1,…N, N is number of prediction instances) |  |  |
| …(other settings and results for Case 3) |  |  |  |
| Fine-tuning case (optional) | Train (setting#A, size/k) |  |  |
| Fine-tune (setting#B, size/k) |  |  |
| Test (setting#B, size/k) |  |  |
| SGCS (1,…N, N is number of prediction instances) |  |  |
| NMSE (1,…N, N is number of prediction instances) |  |  |
| …(other settings and results for Fine-tuning) |  |  |  |
| FFS others |  |  |  |

The intermediate KPI results are in the form of absolute value and the gain over a given benchmark, e.g., in terms of “absolute value (gain over benchmark)”. SGCS is to be expressed in linear domain, while NMSE in dB domain.

***Observations***:

**CSI compression**

From the results for the *generalization verification* of AI/ML based CSI compression *over various deployment scenarios*:

15 sources show that compared to the case 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, it has degraded performance if the model is trained with deployment scenario#A and applied for inference with a different deployment scenario#B.

e.g., deployment scenario#A is UMa, deployment scenario#B is UMi, deployment scenario#A is UMi, deployment scenario#B is UMa, or deployment scenario#A is InH, deployment scenario#B is UMa/UMi.

6 sources observe that if deployment scenario#A and deployment scenario#B are subject to some certain combinations, the degradation is minor.

e.g., deployment scenario#A is UMa, deployment scenario#B is UMi, or deployment scenario#A is UMi, deployment scenario#B is UMa.

6 sources show that generalized performance of the AI/ML model can be achieved, if the training dataset is constructed with data samples subject to multiple deployment scenarios including deployment scenario#A and deployment scenario#B, and the trained AI/ML model applies inference on either deployment scenario#A or deployment scenario#B.

e.g., deployment scenario#A is InH, deployment scenario#B is UMa and/or UMi.

3 sources show that, compared to the case where the AI/ML model is trained on scenario#A and applied for inference on deployment scenario#B, the generalization performance can be improved, if the AI/ML model, after trained on deployment scenario#A, is updated based on a fine-tuned dataset subject to deployment scenario#B, and performs inference on deployment scenario#B.

e.g., deployment scenario#A is InH, deployment scenario#B is UMa or UMi.

From the results for the *generalization 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,

Generalized performance of the AI/ML model can be achieved (0%~5.9% loss) under generalization Case 3 for the inference on either CSI payload size#A or CSI payload size#B, if the training dataset is constructed with data samples subject to multiple CSI payload sizes including CSI payload size#A and CSI payload size#B, and an appropriate scalability solution is performed to scale the dimension of the AI/ML model, shown by 7 sources (6 sources showing 0%~2.2% loss, 3 sources showing 2.35%~5.9% loss). The scalability solution is adopted as follows:

* Pre/post-processing of truncation/padding, adopted by 3 sources, showing 0.2%~5.9% loss.
* Various quantization granularities, adopted by 1 source, showing 1.8%~4.7% loss.
* Adaptation layer in the AL/ML model, adopted by 3 sources, showing 0%~4.05% loss.

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.

**CSI Prediction**

For the AI/ML based CSI prediction,

11 sources show that the AI/ML-based CSI prediction outperforms the benchmark of the nearest historical CSI, wherein

5 sources show the gain of 14% ~ 26.47% using raw channel matrix as input.

2 sources show the gain of 5.64% ~ 9.49% using precoding matrix as input, which is in general worse than using raw channel matrix as input.

Note: spatial consistency is adopted in 1 source and not adopted in 5 sources.

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 is 30km/h.
* The performance metric is SGCS in linear value for layer 1.

## 6.3 Beam Management

### 6.3.1 Evaluation assumptions, methodology and KPIs

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)
* **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 inputed 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 network-sided model.

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:
	+ "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

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:
	+ (FFS) (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[/K] 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, conside the following:

* 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 A: Set B is changed following a set of pre-configured patterns
		- Opt B: Set B is randomly changed among pre-configured patterns
		- Opt C: Set B is randomly changed among Set A beams (pairs)
		- Opt D: 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% outdoor

For 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 (Companies explain how is oppler ) |
| 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

Table 6.3.2-1 presents the performance results.

Table 6.3.2-1: Evaluation results for [BM-Case1 or BM-Case2] without model generalization for [DL Tx beam prediction or Tx-Rx beam pair prediction or Rx beam prediction

|  |  |  |
| --- | --- | --- |
|  | Company A | … |
| Assumptions | Number of [beams/beam pairs] in Set A |  |  |
| Number of [beams/beam pairs] in Set B |  |  |
| Baseline scheme |  |  |
| AI/ML model input/output | Model input |  |  |
| Model output |  |  |
| Data Size | Training  |  |  |
| Testing  |  |  |
| AI/ML model | [Short model description] |  |  |
| Model complexity |  |  |
| Computational complexity |  |  |
| Evaluation results [With AI/ML / baseline] | [Beam prediction accuracy (%)] | [KPI A] |  |  |
| [KPI B]… |  |  |
| [L1-RSRP Diff] | [Average L1-RSRP diff]… |  |  |
| [System performance] | [RS overhead Reduction (%) / RS overhead] |  |  |
| [UCI report] |  |  |
| [UPT]… |  |  |

***Observations***:

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

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 x%~y%, if applicable] in beam prediction accuracy compared to unquantized L1-RSRPs of beams in Set B.

When *Set B is a subset of Set A*, AI/ML can provide good beam prediction performance with less measurement/RS overhead without considering generalization aspects *with the measurements from the best Rx beam* without UE rotation:

(A)With measurements of fixed Set B of beams corresponding to 1/4 of Set A beams:

* evaluation results [from 4 sources] indicate that, AI/ML can achieve [about 70%~80%] beam prediction accuracy of Top-1 DL Tx beam, evaluation results [from 6 sources] indicate that, AI/ML can achieve [about 80%~90%] beam prediction accuracy of Top-1 DL Tx beam, and evaluation results [from 4 sources] show [more than 90%] beam prediction accuracy of Top-1 DL Tx beam.
* evaluation results [from 8 sources] indicate that, AI/ML can achieve [more than 90%] beam prediction accuracy for Top-1 DL Tx beam with 1dB margin.
* evaluation results [from 8 sources] indicate that, AI/ML can achieve [more than 80%] beam prediction accuracy for Top-2 DL Tx beam. The beam prediction accuracy increases with K.
* evaluation results [from 9 sources] indicate that, the average L1-RSRP difference of Top-1 predicted beam can be [below or about 1dB].

(B) With measurements of fixed Set B of beams corresponding to 1/8 of Set A beams:

* evaluation results [from 2 sources] indicate that, AI/ML can achieve [about 50%] beam prediction accuracy of Top-1 DL Tx beam, evaluation results [from 3 sources] show [about 60%~70%] beam prediction accuracy of Top-1 DL Tx beam, and evaluation results [from 2 sources] show [about 70%~80] beam prediction accuracy of Top-1 DL Tx beam.
* evaluation results [from 4 sources] indicate that, AI/ML can achieve [70%-90%] beam prediction accuracy for Top-1 DL Tx beam prediction with 1dB margin
* evaluation results [from 2 sources] indicate that, AI/ML can achieve [about 70%~ 80%] beam prediction accuracy for Top-2 DL Tx beam, and evaluation results [from 4 sources] indicate that, AI/ML can achieve [more than 80%] beam prediction accuracy for Top-2 DL Tx beam. The beam prediction accuracy increases with K.

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.

(C) For the case that Set B of beams is changed among pre-configured patterns, evaluation results [from 4 sources] show that the beam prediction accuracy degrades [no more than 5%] in terms of Top-1 beam prediction accuracy compared to when Set B is fixed across training and inference, where the [one source] used [24] pre-configured patterns and the rest of sources use [4 or 5] patterns; evaluation results [from 1 source] show that the beam prediction accuracy degrades [about 10%] in terms of Top-1 beam prediction accuracy compared to when Set B is fixed across training and inference.

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. The measurements are obtained from the best Rx beam without UE rotation.

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

## 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.

***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 and computational complexity, e.g., FLOPS
	+ 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.
* 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)
	+ 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.
* Other aspects are not excluded.

Companies can evaluate the impact of at least 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 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 6-5. The parameters in the table are applicable to InF-DH at least. If other InF sub-scenario is prioritized in addition to InF-DH, some parameters in Table 6-5 may be updated:

**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 Section 7.5 of TR 38.901 and correlation distance = *dclutter*/2 for InF (Section 7.6.3.1 of TR 38.901)
* the small scale parameters are according to Section 7.6.3.1 of TR 38.901
* the absolute time of arrival is according to Section 7.6.9 of TR 38.901

Baseline evaluation does not incorporate spatially consistent UT/BS mobility modelling (Section 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 to 0.
	+ 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. For a given AI/ML model, the set of TRPs (N’TRP) that provide measurements is fixed.
	+ For Approach 2: 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.
	+ 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 below:



For the evaluation of AI/ML based positioning method, the measurement size and signalling overhead for the model input is 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 evaluations, companies to report assumed sampling period.

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 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:

1. the type of information (e.g., ToA, RSTD, AoD, AoA, LOS/NLOS indicator) to use as model output,
2. soft information vs hard information,
3. 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, unsupervised 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.
* [Whether/how to study the impact of labelling error to label-based model monitoring methods]
* [Whether/how to study the impact of labelling error for AI/ML assisted positioning.]

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.
* 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:

1. 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.
2. 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 6-6 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

If fine-tuning is not evaluated, Table 6.4.2-1 presents the performance results.

Evaluation area shall be included in the evaluations reporting template, assuming the same evaluation area is used for training dataset and test dataset. Note that the baseline evaluation area for InF-DH = 120x60 m. If different evaluation areas are used for training dataset and test dataset, they are marked out separately under “Train” and “Test” instead.

**Table 6.4.2-1: Evaluation results for AI/ML model deployed on [UE or network]-side, [with or without] model generalization, [short model description], UE distribution area = [e.g., 120x60 m, 100x40 m]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Input | Model Output | Label | Clutter parameters | Dataset Size | AI/ML complexity | Horizontal positioning accuracy at CDF=90% (m) |
| Train | Test | Model compl. | Compu compl. | AI/ML |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

If fine-tuning is evaluated, Table 6.4.2-2 presents the performance results.

**Table 6.4.2-2: Evaluation results for AI/ML model deployed on [UE or network]-side, [with or without] model generalization, [short model description], UE distribution area = [e.g., 120x60 m, 100x40 m]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Input | Model Output | Label | Settings (e.g., drops, clutter param, mix) | Dataset Size | AI/ML complexity | Horizontal positioning accuracy at CDF=90% (m) |
| Train | Fine-tune | Test | Train | Fine-tune | Test | Model compl. | Compu compl. | AI/ML |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

***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 [submitted to RAN1#111] 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.

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 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 higher complexity model.

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.

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 [submitted to RAN1#111] 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 [submitted to RAN1#111] 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.

*Generalization*

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%.

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 submitted to RAN1#112bis 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 submitted to RAN1#112bis 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 submitted to RAN1#112bis 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 submitted to RAN1#112bis 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 submitted to RAN1#112bis show the positioning error of (t1, t2)=(50ns, 10ns) is 0.74~0.83 times that of (t1, t2)=(50ns, 50ns).
	+ For the case of (t1, t2)=(50ns, 0ns), evaluation results submitted to RAN1#112bis show the positioning error of (t1, t2)=(50ns, 0ns) is 0.73~0.82 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 submitted to RAN1#112bis 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 submitted to RAN1#112bis 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%.

# 7 Potential Specification Impact Assessment

## 7.1 General observations

[Editor’s note: this section is meant to capture general observations on specification impact considering possibly, different timelines (e.g, short-term vs. long-term)]

## 7.2 Physical layer aspects

In this section, 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.

### 7.2.1 Common framework

### 7.2.2 CSI feedback enhancement

***Items considered for study 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 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 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 considered.

*Intermediate KPI based model monitoring:*

* 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 model*:

* Potential 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 using 3GPP aware mechanism.

*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

*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.
* Signaling 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
* Signaling for triggering the data collection
* 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.

*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.

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/signaling overhead.

Feasibility and methods to support the legacy CSI reporting principles:

* The priority rule regarding CSI collision handling and CSI omission
* Codebook subset restriction
* CSI processing Unit

*Potential specification enhancement on:*

* CSI-RS configurations (not including CSI-RS pattern design enhancements)
* CSI reporting configurations
* CSI report UCI mapping/priority/omission
* CSI processing procedures

### 7.2.3 Beam management

***Items considered for study 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:

* Type1 performance monitoring:
	+ Configuration/Signalling from gNB to UE for measurement and/or reporting
	+ UE may have different operations
		- Option1: UE sends reporting to NW (e.g., for the calculation of performance metric at NW)
		- Option2: 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
* Type2 performance monitoring (UE-side 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
	+ If it is for UE-side model monitoring, UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation
	+ UE reporting of beam measurement(s) based on a set of beams indicated by gNB
	+ Signalling, e.g., RRC-based, L1-based
	+ Note: Performance and UE complexity, power consumption should be considered

*L1 signalling:*

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.

For BM-Case1 and BM-Case2 with a network-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 and BM-Case2 with a UE-side AI/ML model:

* 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)
* Reporting of best beam(s) obtained by measuring beams of a set of indicated by gNB (e.g., Beams from Set A)
* Reporting of 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)

*Data collection:*

At UE side for UE-side AI/ML model:

* UE reporting to NW supported/preferred configurations of DL RS transmission.
* Trigger to initiate 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., signaling aspects related to assistance information (if supported), Reference signals, 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 (if supported)

At NW side:

* Mechanism related to the reporting.
* Additional information for content of the reporting.
* Reporting overhead reduction.
* Signalling/configuration/measurement/report for data collection

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.

*Model Inference related*:

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)

### 7.2.4 Positioning accuracy enhancements

***Items considered for study the necessity, feasibility, potential specification impact***:

*AI/ML model indication[/configuration]*:

* 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

*Signalling, report/feedback*:

* 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
* 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

*Training data generation* for AI/ML based positioning:

* The following options of entity and mechanisms to generate ground truth label are identified:
	+ 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)
* 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

*Training data collection* for AI/ML based positioning:

* Associated information of training data:
	+ Ground truth label at least for model training; report from the label data generation entity
	+ Measurement (corresponding to model input) at least for model training; report from the measurement data generation entity.
	+ Quality indicator for and/or associated with ground truth label and/or measurement at least for model training; 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: 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 training data for model training; report from data generation entity together with training data and/or as LMF assistance signalling.
		- Separate time stamp for measurement and ground truth label, when measurement and ground truth label are generated by different entities.
		- Note: 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
* Assistance signalling and procedure to facilitate generating/collecting training data:
	+ Potential determination of the UE/PRU/TRP which can provide the training data
	+ Configuration of reference signal (for measurement and/or label)
	+ Signalling other than above 2 for data collection, e.g., requested quality of training data

*Model monitoring*:

* 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 network-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)
* If model monitoring does not require ground truth label (or its approximation).
	+ Monitoring metric: e.g., 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.
	+ 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.
	+ 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

*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 signaling 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
* Assistance signalling and procedure to facilitate model inference for both UE-side and Network-side model
	+ RS configurations

*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.

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

## 7.3 Protocol aspects

In this section, aspects related to, e.g., capability indication, configuration and control procedures (training/inference), and management of data and AI/ML model, per RAN1 input, are considered.

In addition, collaboration level specific specification impact per use case is documented.

 *Editor’s note: The text above will be updated based on the progress of the study/discussion.*

### 7.3.1 Common framework

#### 7.3.1.1 Model and Functionality Identification

Model is identified by a model ID. To manage or control AI/ML models some metadata about them may be needed. In this regard, and similar to what is captured in clause 4.2, from a RAN2 perspective, this meta information could come in the form of a model ID which can be used to identify an AI/ML model or a set of AI/ML models. RAN2 assumes that a model ID is globally unique, so that it allows for proper model training, model validation, and model testing procedures.

Note: Details of model training, validation and testing are out of RAN2 scope.

*Editor’s note: It is still FFS in RAN2 how to define (or eventually achieve) uniqueness of model IDs.*

*Editor’s note: It is still FFS in RAN2 which other metadata can be used to control or manage AI/ML models (e.g., whether to include vendor information, applicable conditions of models, model performance indicators, etc...).*

According to the functional framework in Figure 4.4-1, for a model-ID-based LCM, a model ID can be used within functions (e.g., Inference, Model Storage, Model Training) and for different data/information flows. For example, a model ID could eventually be associated to a “Management Instruction” (e.g., selection/(de)activation/switching), or linked to the “Model Transfer/Delivery” information.

#### 7.3.1.2 Data collection

Data collection plays a crucial role in enabling the different use cases. Within RAN2, extensive discussions have taken place to define the best approaches for collecting data to support model inference, monitoring, and training.

To provide a comprehensive overview of different available data collection methods, RAN2 has endorsed Table 7.3.1.2-1 which describes key indicators of each to be considered, to support model training at the network. An extension to the existing table has been agreed upon. This enhanced table includes three additional columns: Inference, Monitoring, and Training. These columns serve as valuable resources, detailing the applicability of data collection techniques to the LCM purpose and specific use cases.

RAN2 acknowledges that certain scenarios may not require additional specification efforts for data collection. For instance, when model inference is performed on UE-sided models, the required input data is readily available within the UE itself. Similarly, UE-side monitoring of UE-sided models can be independently conducted, leveraging the performance metrics readily accessible within the UE. In such cases, the existing data sources suffice, reducing the need for additional RAN2 specifications.

Considering the importance of latency in data collection, RAN2 has assumed certain requirements concerning the timely availability of data. While offline model training for all model sidedness scenarios (i.e., UE-sided, NW-sided, and two-sided model) appear not to impose any specific latency requirements, situations where model inference or monitoring relies on data from other entities necessitate meeting latency constraints for efficient operations.

Furthermore, RAN2 has primarily focused on the RRC\_CONNECTED state in its analysis and selection of data collection frameworks. By prioritizing this state, which should cover both data generation and reporting, RAN2 aims to ensure a robust foundation for effective data collection.

In scenarios where data generation and termination entities are deployed separately, RAN2 has outlined assumptions specific to CSI enhancement and beam management (see clauses 7.3.2 and 7.3.3, respectively), as well as positioning enhancement use cases (see clause 7.3.4). These assumptions provide guidance on the generation and termination of data for different model-related activities, facilitating seamless communication and collaboration between entities involved.]

Table 7.3.1.2-1. Existing data collection methods identified

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Involved Network entity** | **RRC state to generate data** | **Max payload size per reporting\*** | **Contents to be collected** | **End-to-End report latency\*\*** | **Report type** | **Security and Privacy** |
| Logged MDT | TCE/OAM(It can be utilized by gNB) | RRC\_IDLE/RRRC\_INACTIVE | <9kbyte | L3 cell/beam measurements, location info, sensor info,timing info | 1. Procedure latency\*\*\*:
	* Latency to enter CONNECTED state
	* Latency to receive gNB request signaling (~20ms)
2. Air interface signaling latency\*\*\*\*:
	* ~20ms (RRC)
3. Other latency:
	* Forwarding latency between gNB and TCE
 | Upon gNB request after entering RRC\_CONNECTED | AS security via RRC message,Privacy via user consent  |
| Immediate MDT | TCE/OAM(It can be utilized by gNB) | RRC\_CONNECTED | <9kbyte | L3 cell/beam measurements, location info, sensor info | 1. Procedure latency:
	* Report interval:
		+ l20ms~30min for periodic report
		+ TTT for event triggered report
2. Air interface signaling latency:
	* ~20ms (RRC)
3. Other latency:
	* Forwarding latency between gNB and TCE
 | Event triggered report,Periodic reporting | AS security via RRC message,Privacy via user consent |
| L3 measurements | gNB | RRC\_CONNECTED | <9kbyte | L3 cell/beam measurements | 1. Procedure latency:
	* Report interval:
		+ l20ms~30min for periodic report
		+ TTT for event triggered report
2. Air interface signaling latency:
	* 20ms (RRC)
 | Event triggered report,Periodic reporting | AS security via RRC message. |
| L1 measurement (CSI reporting) | gNB | RRC\_CONNECTED | <1706bit in PUCCH, <3840bit in PUSCH | L1 CSI measurement | 1. Procedure latency:
	* Report interval:
		+ 4-320 slot for periodic report and semi-persistent report
		+ 0-32 slot after reception of DCI for aperiodic report
2. Air interface signaling latency:
	* 1 TTI (PUCCH)
 | Aperiodic report,Semi-persistent report,Periodic report | No AS security |
| UAI | gNB | RRC\_CONNECTED | <9kbyte | Assistance information to show UE preference | 1. Procedure latency:
	* Upon generation of UE's preference
2. Air interface signaling latency:
	* ~20ms (RRC)
 | Up to UE implementation when to report | AS security via RRC message |
| Early measurements | gNB | RRC\_IDLE/RRC\_INACTIVE | <9kbyte | L3 cell/beam measurements | 1. Procedure latency:
	* Latency to enter CONNECTED state
	* Latency to receive gNB request signaling (~20ms)
2. Air interface signaling latency:
	* ~20ms (RRC)
 | Upon gNB request after entering RRC\_CONNECTED | AS security via RRC message |
| LPP | LMF | RRC\_CONNECTED | <9kbyte | Location info | 1. Procedure latency:
	* Latency to get upper layer trigger (for UE triggered)
	* Or latency to receive NW request message (~20ms)
2. Air interface signaling latency:
	* ~20ms (RRC)
3. Other latency:
	* Forwarding latency between gNB and LMF
 | UE-triggered,NW-triggered | AS security via RRC message |

Note:
\* 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.

#### 7.3.1.3 Model Transfer/Delivery

*Editor’s note: Further discussion is needed in RAN2 to complete this clause.*

To analyse the feasibility and benefits of model transfer/delivery, the following solutions are considered from a RAN2 perspective:

* Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.
* Solution 2a: CN (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: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.
* Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.
* Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).

The solutions map to use cases according to what is depicted in Table 7.3.1.3-1.

Table 7.3.1.3-1 Relations between model transfer/delivery solutions and use cases

|  |  |
| --- | --- |
| **Solutions** | **Applicable use cases** |
| Solution 1a, 1b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 1a and 1b. |
| Solution 2a, 2b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 2a and 2b. |
| Solution 3a, 3b | Positioning accuracy enhancement |
| Solution 4 | CSI feedback enhancementBeam managementPositioning accuracy enhancement |

*Editor’s note: The solution-to-use case relation is work in progress.*

#### 7.3.1.4 UE Capability Reporting

*Editor’s note: It is still FFS in RAN2 if for UE capability for AIML methods we use the UE capability mechanisms as defined for RRC reported and LPP reported capabilities.*

#### 7.3.1.5 Applicability reporting of functionalities and models

*Editor’s note: From what is discussed in clause 4.2, further RAN2-centric details/options could be included in this part.*

### 7.3.2 CSI feedback enhancement

RAN2 has identified a set of objectives for the two-sided CSI compression use case. Firstly, to ensure that the UE-part and gNB-part of the models are configured and setup according to their applicable scenarios and configuration. Secondly, to ensure that models match accurately, ensuring that the CSI encoder used at the UE corresponds to the CSI decoder 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 all CSI compression use cases, the selection, (de)activation, switching, and fallback of models or 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-sided versus network-sided models.

For data collection analysis, various scenarios unfold from a RAN2 perspective when the data generation and termination entities are at different entities. For instance, for:

* Model Training:

	+ Training data can be generated by either the UE or the gNB, depending on specific requirements, while the termination point for training data includes the gNB, OAM, or Over-The-Top (OTT) server.
* Inference:

	+ For network-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-sided model inference, the gNB can generate input data or assistance information while the termination point for this data lies within the UE, where the inference process is performed.
* Monitoring (i.e., within the Management function):

	+ For monitoring at the network side of UE-sided model, the UE can generate performance metrics while the termination point for these metrics is the gNB.

### 7.3.3 Beam management

As it is for the CSI use cases, 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 data collection analysis, various scenarios unfold from a RAN2 perspective when the data generation and termination entities are at different entities. In this case, the same list as the one depicted in clause 7.3.2 applies for Model Training, Inference and Monitoring.

### 7.3.4 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-sided versus network-sided models.

For data collection analysis, various scenarios unfold from a RAN2 perspective when the data generation and termination entities are at different entities. For instance, for:

* Model Training:

	+ Training data can be generated by either the UE or the gNB, depending on specific requirements, while the termination point for training data includes the LMF, or OTT server.

*Editor´s note: RAN2 to discuss if the gNB should be added as termination point to address the following RAN1 scenario: “Case 3a: NG-RAN node assisted positioning with gNB-sided model, AI/ML assisted positioning”.*

* Inference:

	+ For network-sided model inference, the UE or the gNB can generate the necessary input data while the termination point for this input data could lie within the LMF or an OTT server, where the inference process is performed.

*Editor´s note: RAN2 to discuss if the gNB should be added as termination point to address the following RAN1 scenario: “Case 3a: NG-RAN node assisted positioning with gNB-sided model, AI/ML assisted positioning”.*

* + For UE-sided model inference, the gNB or LMF can generate input data or assistance information while the termination point for this data lies within the UE, where the inference process is performed.
* Monitoring (i.e., within the Management function):
	+ For monitoring at the network side of UE-sided model, the UE or the gNB can generate performance metrics while the termination point for these metrics is the LMF.

## 7.4 Interoperability and testability aspects

In this section, 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.

### 7.4.1 Common framework

### 7.4.2 CSI feedback enhancement

### 7.4.3 Beam management

### 7.4.4 Positioning accuracy enhancements

# 8 Conclusions

[Editor’s note: conclusions may include recommendations for subsequent WI(s).]

Annex <X> :
Change history

Use style "Heading 8" in TSs and "Heading 9" in TRs. Do not use "informative" in the title in TRs.

This is the last annex for TS/TSs which details the change history using the following table.
This table is to be used for recording progress during the WG drafting process till TSG approval of this TS/TR.
For TRs under change control, use one line per approved Change Request
Date: use format YYYY-MM
CR: four digits, leading zeros as necessary
Rev: blank, or number (max two digits)
Cat: use one of the letters A, B, C, D, F
Subject/Comment: for TSs under change control, include full text of the subject field of the Change Request cover
New vers: use format [n]n.[n]n.[n]n

|  |
| --- |
| **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 | R1-2306170 |  |  |  | RAN1 agreement up to and including RAN1#112bis-e  | 0.1.0 |
|  |  |  |  |  |  |  |  |

Annex <Y>:
List of RAN2 Agreements

Below the main agreements, observations and assumptions captured in the different RAN2 meeting discussions. Those highlighted are captured in the TP above.

**RAN2#119bis-e (October 10 – 19, 2022)**

Some initial Assumptions on the work:

- Assume that RAN2’s work can be somewhat split: A) use-case-centric configuration, signalling and control procedures, B) management of data and AI/ML models (where part of discussion may overlap between use cases).

- Assume that e.g. for the management of data and AI/ML models, RAN2 could start by focusing on data collection, model transfer, model update, model monitoring and model selection/(de)activation/switching/fallback (to the extent needed), whether UE capabilities has a role in this.

- Chair assumes that we will input on various aspects when the time is right, and e.g. postpone things that obviously need R1 decisions, but there could be some rare exception.

**AIML methods**

* Assume that R2 will reuse terminology defined by R1 to the extent possible/reasonable
* Observation: the collaboration levels definitions doesn’t really clarify what is required, more work is needed
* R2 assumes that for the existing (under discussion) AI/ML use cases, proprietary models may be supported and/or open format may be supported (and maybe RAN2 doesn’t have to further elaborate on this assumption).
* R2 assumes that from Management or Control point of view mainly some meta info about a model may need to be known, details FFS.
* R2 assumes that a model is identified by a model ID. Its usage is FFS.
* General FFS: AIML Model delivery to the UE may have different options, Control-plane (multiple subvariants), User Plane, can be discussed case by case.

**RAN2#120 (Toulouse, France, November 14 – 18, 2022)**

**AIML methods**

* R2 assumes that model ID can be used to identify which AI/ML model is being used in LCM including model delivery.
* R2 assumes that model ID can be used to identify a model (or models) during model selection/activation/deactivation/switching (can later align with R1 if needed).
* For model transfer/delivery for AI/ML models (for the target use cases of this SI), RAN2 to study CP-based, UP-based solutions

**Use case specific aspects**

* RAN2 scope includes procedures, protocols, and signaling for two-sided CSI use case(s), e.g.
1. Ensuring UE and gNB side models are configured / applied based on their applicable configurations / scenarios.
2. Ensuring that models are matched properly at both UE and gNB sides, i.e., when a CSI encoder is used at the UE corresponding CSI decoder is used at the gNB
3. Achieving simultaneous (de)activation and switching of the two-sided model

**RAN2#121 (Athens, Greece, February 27 – March 3, 2023)**

**AIML methods**

Data Collection

*Proposal 1 RAN2 to simultaneously focus on studying data collection solutions for both NW- and UE-sided AIML models, including assistance signalling and (dataset) reporting from the concerning entity.*

*Proposal 2 Study RAN2 implications of data collection for all concerning LCM purpose, e.g., model training/monitoring/selection/update/inference/etc.*

*Proposal 3 RAN2 to separately analyse the data collection requirements and solutions for the different LCM purposes. FFS if general frameworks/solutions could be adopted.*

*Proposal 4 Wait for RAN1 requirements before discussing specific data collection solutions for use cases and for the related (LCM) procedures. In the meantime, RAN2 can summarize the implementation of existing frameworks while focusing on different performance metrics.*

*Proposal 5 When summarizing the different data collection frameworks, RAN2 can start by considering the following metrics: a) the content of the data, b) the data size, c) latency and periodicity, d) signalling, entities involved, and configuration aspects. FFS on how to handle security/privacy.*

*Proposal 6 Consider the following existing frameworks as starting points to be considered for data collection: SON & MDT, UE assistance information, RRM measurement reports, CSI reporting framework, LPP Provide location information. FFS whether other frameworks should be discussed.*

*Proposal 7 Upon receiving specific (RAN1) requirements, RAN2 to decide whether the existing frameworks can be reused/extended, or whether a new framework is required.*

*Proposal 8 For data collection, RAN2 will simply keep progressing and will inform of concerning agreements to RAN1 when necessary.*

* P1-P8 are loosely endorsed with the understanding that we can also go beyond, e.g. analyse other methods.

Rapporteur’s Note: The following agreement is referring to *[R2-2300708](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip)*:

* The table in this doc is endorsed as starting point

Rapporteur’s Note: The table in *[R2-2300708](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip)* (see agreement just above) led to a further iteration in [R2-2302286](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302286.zip) and the following set of agreements:

* Endorse the table as a starting point (e.g. can add more columns if needed later, modify, add rows etc). Content shall be interpreted as current content.
* Chair: There is significant support to aim for evaluating the data collection methods per LCM purpose

Model Transfer

* We Use the wording “model transfer/delivery”
* model delivery that serves the use cases in the SI is within RAN2 scope, regardless other aspects.
* Agreed:

Aim to at least analyze the feasibility and benefits of model/transfer solutions based on the following:

Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

Solution 2a: CN (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: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).

**Table: relations between solutions and use cases**

|  |  |
| --- | --- |
| **Solutions** | **Applicable use cases** |
| Solution 1a, 1b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 1a and 1b. |
| Solution 2a, 2b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 2a and 2b. |
| Solution 3a, 3b | Positioning accuracy enhancement |
| Solution 4 | CSI feedback enhancementBeam managementPositioning accuracy enhancement |

Note: the solutions use case relation is preliminary (work in progress), and the purpose is to have better understanding on what to further analyse

Chair think that in general, we may need to understand what issues are expected, e.g. Loosely Expect that time/latency from trigger to get a new model and until is downloaded and operational may be an issue, expect some other issue (in certain circumstances) and so on …

Rapporteur’s Note: The following agreement is referring to [R2-2302268](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302268.zip):

* The table can serve as starting point for continued discussion (but contains some parts that seems non consensus, e.g. delta configuration).

Model ID and UE cap

* RAN2 assumes that Model ID is unique “globally”, e.g. in order to manage test certification each retrained version need to be identified.

**General**

* R2 may consider including the existing EVEX framework for this SI, FFS exactly what this means, can discuss next meeting.

**RAN2#121bis-e (April 17 – 26, 2023)**

**AIML methods**

* R2 will deprioritize aspects of on-line/real-time training for the whole SI (unless R1 identifies that it is needed for one of the studied use cases).

Architecture General

* FFS if For UE capability for AIML methods we use the UE capability mechanisms as defined for RRC reported and LPP reported capabilities.
* For the CSI compression and beam management use cases, model/function selection/(de)activation/switching/fallback can be UE-initiated or gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
* For the positioning use case, model/function selection/(de)activation/switching/fallback can be UE-initiated or LMF-/ gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
* R2 assumes that Information such as FFS:vendor info, applicable conditions, model performance indicators, etc. may be required for model management and control, and should, as a starting point, be part of meta information.
* The general AI/ML framework consist of, (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage.

Chair: the following was almost agreed (leave it FFS for now): AI/ML functional architecture in Figure 1 in [R2-2303674](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2303674.zip) is the baseline with the modification that Performance Monitoring is changed to Model Mgmt / Performance Monitoring. It is noted that the exact interactions may need some modification depending on how each piece of functionality is specified**.**

* Model ID can be used to identify model or models for the following LCM purposes:

model selection/activation/deactivation/switching (or identification, if that will be supported as a separate step).

(e.g. for so called “model ID based LCM”)

* If model transfer/delivery is supported, model ID can be used for model transfer/delivery LCM purpose.
* How to achieve globality of the Model ID is FFS.

Initial discussion in RAN2: the following global unique model ID definition directions can be considered as a starting point:

Direction1: Pre-defined/hard-coded global unique model ID

Direction3: Assigned global unique model ID via specific ID management node.

Note: Other global unique model ID definition is not precluded.

Model ID structure, if any, is FFS

Chair: companies can also consider the remaining proposals and proposed open issues for later discussions.

Rapporteur’s Note: The chair’s observation above is referring to proposals and open issues in [R2-2304195](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2304195.zip).

Data Collection

* Extend the previously endorsed table with 3 columns: Inference, Monitoring and Training, and explain in free text the applicability of the data collection method to the LCM purpose and the use case(s).
* Observation: RAN2 may need to consider enhancements for AIML to existing functionality for data collection, e.g. for timing control (e.g. for MDT/RRM).

Rapporteur’s Note: The following set of agreements relate to [R2-2304541](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2304541.zip).

* P1: RAN2 to understand/determine/capture requirements of data collection for the LCM functionalities and document the results. FFS on the exact presentation format. Expect RAN1 to provide some related information.
* P2: RAN2 to capture the analysis (see P1 above) separately for the use-cases, i.e., CSI feedback enhancement, beam management and positioning enhancement. FFS how we do the formatting/presentation of the results.
* P3: Study the applicability (and limitations) of each identified data collection framework for each of the identified LCM purposes, i.e., inference, monitoring and (offline) training. FFS how we do the formatting/presentation of the results.
* P4: With more progress on architectural discussion, consider the suitability of each identified data collection framework for the termination points and mapping with the location of LCM purposes/functions (inference, monitoring, (offline) training)

- Model sidedness (UE side, NW side, two sided) FFS

- Use case mapping FFS

* P5: RAN2 to modify the previously endorsed table by adding 3 additional columns: inference; monitoring and (offline) training. Whether to, and how to further restructure the table is FFS.

Rapporteur’s Note: The following chair comments regarding EVEX where based from online discussion on [R2-2302954](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2302954.zip).

Chair: There is some support to add EVEX as an option, but there is a lot of concerns. Majority of companies seems to have concerns.

Chair: Maybe the vivo proposal was too wide: Proposal: Add EVEX (or modified EVEX if needed) as one potential option for collection of data for training for UE side models.

- Huawei, ZTE, OPPO, CMCC, Ericsson and Apple object

**RAN2#122 (Incheon, Republic of Korea, May 22 – 26, 2023)**

Functional Arch

* Intention is to cover functional arch in general, e.g. covering both be model based and/or functionality based LCM
* “Model Storage” in the figure is only intended as a reference point (if any) for protocol terminations etc for model transfer/delivery etc. It is not intended to limit where models are actually stored. Add a note for this.
* Remove “Model” in Model Managemt and Model Inference and for the actions/the arrow form Management to Inference (to reduce the risk for misunderstanding).
* Management may be model based management, or functionality based management. Add a mote for this.
* With the modifications above Figure 2 from [R2-2305327](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_122/Docs//R2-2305327.zip) is agreed

Data Collection

Rapporteur’s Note: The following set of agreements relate to [R2-2306783](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_122/Docs//R2-2306783.zip)

* P1a: For the LS to RAN1 on data collection requirement, inform RAN1 that the reply should be per use case and per LCM purpose (i.e., Model training, inference and monitoring), and LCM sidedness should also be considered.
* RAN 2 assumes that for the data collection in some scenarios (e.g., internal data up to implementation or the existing data are enough), possibly no RAN2 specification effort is needed in some scenarios, e.g. (not exhaustive):

- For model inference of UE-sided model, input data for model inference is available inside the UE.

- For UE-side (real time) monitoring of UE-sided model, performance metrics are available inside the UE. UE can independently monitor a model's performance without any data input from NW.

* P2a: LS to ask RAN1 to provide the required data content per use case and per LCM purpose, when available, and to what extent said data would / should be specified (in detail).
* P2b: LS to ask RAN1 about the reporting type (e.g., periodic, event triggered, other) of the identified data content.
* P3: LS to ask RAN1 about the typical size (value or value range) of the identified data content.
* P4a: For the latency requirement of data collection, RAN2 assumes:

- 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 model monitoring, when required monitoring data (e.g., performance metric) comes from the other entities, there is a latency requirement for data collection.

* P4b: LS to RAN1 to confirm the WA (in P4a) on the latency requirement, and ask RAN1 about the typical latency requirement (value or value range) to transfer the identified data content.
* P6a: RAN2 assumes 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 on the non-connected state can be revisited when needed.
* P6b: LS to RAN1 to confirm the WA (in P6a) on RRC state of data collection.
* P5a: For the data generation entity and termination entity deployed at different entities, RAN2 assumes:

For CSI enhancement and beam management use cases:

- For model training, training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server.

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

- For UE-side model inference, input data/assistance information can be generated by gNB and terminated at UE.

- For model monitoring at NW side, performance metrics can be generated by UE and terminated at gNB.

For positioning enhancement use case:

- For model training, training data can be generated by UE/gNB and terminated at LMF/OTT server.

- For NW-sided model inference, input data can be generated by UE/gNB and terminated at LMF and/or gNB.

- For UE-side model inference, input data/assistance information can be generated by LMF/gNB and terminated at the UE.

- For model monitoring at NW side, performance metrics can be generated by UE/gNB and terminated at LMF.

* P5b: LS to RAN1 to confirm the WA (in P5a) on the generation entity and termination entity of the identified data content and ask for supplement, if any.

Rapporteur’s Note: Regarding the LS out to RAN1 on Data Collection Requirements and Assumptions:

* Approved in [R2-2306906](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_122/Docs//R2-2306906.zip)