**3GPP TSG-RAN WG4 Meeting # 108bis R4-2317258**

**Xiamen, China, October 09 – October 13, 2023**

**Agenda item:** 5.21.4

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [108bis][135] FS\_NR\_AIML\_air

**Document for:** Information

# Introduction

This is the summary thread for issues related to NR AI/ML study in RAN4. A WF summarizing online agreements and other issues proposed for agreement was discussed the previous meeting in R4-2314740 but not agreed. However, most companies have provided inputs in this meeting based on this WF. This summary is organized in 3 high level topics and contains several sub-topics for discussion.

# Topic #1: General aspects and TR

This section contains the sub-topics regarding general issues and proposed TR updates

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2315065**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315065.zip) | CAICT | **Proposal 1:** **Latency requirements of data collection for model inference and monitoring should be considered and discussed per use case, subject to the output from RAN1/2.****Proposal 2: Accuracy requirements of data collection with legacy measurements could follow legacy measurements requirements. Potential enhancements to legacy requirements and accuracy requirements on new data types/measurements could be discussed per use case and considered in WI phase.****Proposal 3: Whether/How to define the requirements of generalization performance could be discussed per use case. The progress of RAN1 on generalization verification can be taken as a reference to define generalization tests in RAN4.****Proposal 4: Support option 3 (Option 1 and Option 2 depending on the test cases) of issue 1-6 (RAN4 testing goals) in R18 study phase and following R19 normative phase.** |
| [**R4-2315103**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315103.zip) | CATT | **Proposal 1: RAN4 does not define accuracy requirements for data collection unless RAN4 is requested to do so for some procedures by other WGs.****Proposal 2: RAN4 to define delay requirements for data collection when data are transferred between different entities. Details are FFS.** **Proposal 3: The test scenarios should be limited to the scenarios similar to the scenario the AI/ML models are trained for, e.g. InF vs. office room.** **Proposal 4: The test results of the unseen inputs from a similar scenario directly reflect the robustness of AI/ML models.****Proposal 5: Fine-retuning of model weights could be considered, if necessary, to maintain the test performance.** **Proposal 6: RAN4 should focus on the UE performance other than AI/ML model performance in the tests.****Proposal 7: RAN4 to use only one model with model update (i.e. fine retuning weights) for tests in dynamic changing environment as a starting point at current stage. Multiple-model option can be FFS.** **Proposal 8: RAN4 to focus on whether the performance gain of AI/ML model can be achieved.** |
| [**R4-2315309**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315309.zip) | Qualcomm, Inc. | **Proposal 1: If requirements are needed, the following principles can apply to study how to derive the UE-side performance metric estimation for monitoring requirements that can accommodate different types of proper UE implementations*** **The accuracy requirement should verify whether the performance metric reported by the UE reflects the two-sided model performance, e.g., by verifying if the reported performance metric has enough correlation with the observed performance (e.g., throughput).**

**Proposal 2: RAN4 considers the propagation conditions used in legacy tests as baseline for AI/ML model testing. Modifications on channel model or introduction of new channel models from RAN1 evaluation setup in R18 SI can be considered for specific use cases if the necessity is recognized by RAN4.** **Proposal 3: Generalization and robustness test scope should focus on the AI/ML model inference performance without LCM procedures under different propagation conditions, and the LCM related procedures can be tested separately.****Proposal 4: If RAN4 decides to test the generalization and robustness of AI/ML model inference performance without LCM procedures, the following options can be considered:*** **Option 1: RAN4 defines multiple tests with different propagation conditions. In each of the test, TE signals the same model ID, and therefore the same AI/ML model is tested under different propagation conditions to verify it’s generalizability and robustness.**
* **Option 2:** **RAN4 defines one test and changing different propagation conditions within the test. Therefore the same AI/ML model is tested under different propagation conditions to verify it’s generalizability and robustness.**

**We propose to consider option 1 as the baseline option, and if any agreed performance requirements can’t be verified by option 1, but option 2 or other options can cover those requirements, RAN4 can revisit it.** **Proposal 5: TR draft can capture the studies/agreements in the following issues based on the current progress*** **General aspect of complexity**
* **KPI for the requirement in different use cases**
* **Two sided model: test model options**

**Proposal 6: RAN4 consider the following block diagram to capture the listed open issues studied in AI/ML RAN4 interoperability and testability**Test configurationData generator AI/ML modelModel control/ signalingAI/ML modelPerformance monitoringDUTData collectionTE* **Inference tests: under a static configuration, TE verifies the performance of AI/ML model based on the output it produces with possible post processing by TE’s reference AI/ML model, if two-sided model is considered.**
* **Model management related tests: DUT executes the actions triggered by model control or associated signaling, and TE verifies the reaction on the DUT AI/ML model against the defined requirement.**
* **Two sided model: TE has AI/ML models connected with DUT AI/ML models, and the AI/ML models in the TE are part of the test.**
* **Data collection: RAN4 can collect (emulated) data from DUT AI/ML model output and/or data generator in TE.**

**Note that some of the reporting/monitoring mechanisms are subject to RAN1 agreed procedures and not necessarily feasible/required for RAN4 discussion.** |
| [**R4-2315351**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315351.zip) | CAICT. | TR Updated proposal, to be discussed |
| [**R4-2315598**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315598.zip) | NTT DOCOMO, INC. | **Observation 1: AI/ML model structure/parameters highly depends on implementation, and it is difficult to specify them as a standard.****Proposal 1: RAN4 should focus on following points under arbitral AI/ML model structure/parameters assumption*** **Appropriate AI model selection for each test case**
* **Metric definition and evaluation method for each use case**
* **Metric definition and monitoring method for LCM**

**Proposal 2: AI model in TE for 2-sided model test should be up to TE vendor implementation****Observation 2: Considering the real situation, UE cannot know the detail of AI model of gNB and vice versa****Proposal 3: Test results variation between TEs should be absorbed by generalization performance****Observation 3: No existing test methodology can be applied to evaluate generalization performance****Proposal 4: RAN4 should introduce new test methodology to evaluate generalization performance, for example following two methodologies:**1. **Fuzzy side condition test concept**

**The concept is that side condition has certain variation, e.g. each parameters are defined as probability function, but performance metric is fixed. This fuzzy side condition may evaluate generalization performance.**1. **Morphing test concept**

**The concept is that test case is morphing from the start condition to the end condition continuously during certain evaluation time. This morphing test may evaluate generalization performance.**Proposal 5: Companies are invited to provide feasibility view of these new concepts |
| [**R4-2315726**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315726.zip) | vivo | ***Observation 1: The AI/ML complexity (model complexity and computational complexity) could be considered when specify reference model for defining requirements and discussed case by case.******Observation 2: Post deployment performance may be verified by model monitoring.******Proposal 1: Accuracy requirements for input data collection need to be considered.******Proposal 2: RAN4 should study latency requirements for data collection of model monitoring, at least for positioning and CSI compression use cases.******Proposal 3: RAN4 testing goal for AI/ML is to verify whether a specific AI/ML model can be conducted in a proper way and whether the performance gain of AI/ML model can be achieved for specified scenario/configuration.******Proposal 4: RAN4 generalization testing goal, for both one-sided and two-sided model, for AI/ML is to verify whether the performance gain of AI/ML model can be achieved for various scenarios/configurations.******Proposal 5: The feasibility of generalization test by using dynamic scenarios/configurations needs further study by considering the following aspects.**** ***Testing method***
	+ ***TE changes the scenario/channel in turn and covers all the required N scenarios/channels to be tested***
	+ ***TE chooses a subset of the N scenarios/channels as the scenarios/channels under test based on the certain rules, and changes the scenario/channel randomly and completes the test within the requirement time T.***
* ***Test dataset generation***
	+ ***Stationary statistical channel modelling method***
	+ ***Non-stationary channel modelling method***
	+ ***Field channel measurement***
	+ ***Deterministic channel modelling (e.g., Ray Tracing)***
 |
| [**R4-2315753**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315753.zip) | ZTE Corporation | **Observation 2: For model inference, the MAX total latency values will be .****Proposal 1:** **RAN4 shall define the latency requirements based on RAN2’s agreements and the MAX total latency requirements can be:** where, **t1** denotes the propagation delay from gNB to send the related RS to UE;**t2** denotes the measurement time for RS;**t3** denotes the uncertainty time;**t4** denotes the propagation delay from UE to report the measurement result to gNB (as the NW-sided model inference input);**N** denotes the number of selected samples.In theory, t1=t4 the propagation delay shall be the same which equals the distance (d) between gNB and UE divides the speed of light (c):**Observation 3: The ground truth shall be mapped to the input and compared to measurement results to guarantee the data quality.****Observation 4: If model is sent to the field which the radio conditions and other obstacles ( e.g. NLOS) are totally different and complicated than the test environment, the model performance will be degradation and also the model generalization is not beneficial to RAN4 testing goals.****Observation 5: What is the boundary of the degradation and the tolerant margin can be also defined in order to judge the performance will be degraded or not.****Proposal 2: RAN4 shall consider and study the tolerant margin to judge the performance degradation.****Observation 6: The static scenarios refers to the simplification of data features, with fewer data features, and may not even be representative.** |
| [**R4-2315994**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315994.zip) | Huawei,HiSilicon | ***Proposal 1:*** Legacy requirements for existing use in RAN4 may not be applicable when define AI/ML performance requirements, if the effect of operations from the opposite side is not eliminated or well controlled.***Proposal 2:*** RAN4 AI/ML testing goal is identified from the following options.* Option 1: The testing goal is to verify whether a specific AI/ML model can be conducted in a proper way.
	+ FFS how to determine the specific AI/ML model.
	+ FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)
* Option 2: The testing goal is to verify whether the performance gain of AI/ML model can be achieved for a specific scenario/configuration.
	+ FFS how to determine the specific scenario/configuration, taking account of ensuring the consistency between testing dataset and training dataset.

***Proposal 3:*** Table for RAN4 testing goal when model under test is transferred from the opposite side with and w/o air interface signaling.

|  |  |  |
| --- | --- | --- |
|  | Testing goal (Option 1): verify that the model is properly conducted | Testing goal (Option 2): verify the performance of the model |
| Model under test is transferred from the opposite side with air-interface signaling | √ | - |
| Model under test is delivered w/o air interface signaling  | - | √ |

Note1: Wait RAN1/2 to study whether to specify model transfer with air interface signaling or not.Note2: Whether it is testable or not for selected testing goal is a separate discussion. ***Proposal 3:*** Whether to consider generalization verification needs to wait RAN1/2 progress. If model transfer over the air interface signaling is not specified and generalization is testable after RAN4 studying, then generalization verification reuses legacy RAN4 test, where different requirements may be considered in different scenario/configuration, separately. |
| [**R4-2316228**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316228.zip) | OPPO | **Proposal 1: For RAN4 AI/ML performance requirements** **and tests, following options should be considered,** **- Option 1: Requirements/tests are defined on a feature level, not for a specific solution/model** **- Option 2: Requirements/tests are defined for a specific model** **- Option 3: Requirements/tests are defined for a generalized performance** **- Option 4: Requirements/tests are defined for a scenario-based performance****Proposal 2: Regarding the AI/ML life cycle management impacts and performance monitoring, mechanisms to avoid the interference of random effects on the evaluation results should be studied, including** **- Multi-sample involved performance evaluation to obtain a stable evaluation result** **- Multi-user involved performance evaluation for a cell level feature/model**  |
| [**R4-2316456**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316456.zip) | Ericsson | **Proposal 1**: Send the approved TP to RAN1 in the LS as attachment requesting RAN1 to incorporate the TP in TR38.843.Addition agreed in previous meeting:**Test encoder/decoder for TE**: AI/ML model for UE encoder/gNB decoder implemented by TE. |
| [**R4-2316467**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316467.zip) | MediaTek inc. | ***Observation 1***: **If the testing goal of generalization is to verify inference performance for some specific AI/ML model, the AI/ML model cannot be switched/updated during the generalization test.*****Observation 2*: If it allows UE to perform model switching/update to adapt the change of scenarios/configurations, it is difficult to tell the test failure is due to model switching/update or model inference.*****Observation 3*: Not all use cases considered in RAN1 are good for generalization. Take CSI prediction as an example, the results from RAN1 show that AI/ML models can attain good performance if the UE speed of the training data is the same as that of the testing data, while the performance is worse if the UE speed of the training data is different from that of the testing data. Besides, the AI/ML model trained by the mixed dataset shows moderate performance on each scenario of different UE speeds.*****Observation 4*: There are lots of different scenarios can be taken into consideration for generalization, for example, generalization over different RBs, generalization over different UE speed, generalization over different deployment.** ***Proposal 1*: If the testing goal for generalization is to verify the inference performance of some specific AI/ML model with non-stationary scenarios/configurations. RAN4 should discuss*** **how to prevent model switching/update during the test**
* **the generalization capability for the considered use case**
* **which scenarios used in defining generalization requirement**

***Proposal 2*: Further discuss how many states of AI/ML model may have and what status change expected in each LCM procedure.*****Proposal 3*: Not to discuss whether and how to define requirements for Model/Functionality monitoring in R18 SI stage anymore.*****Proposal 4*: Discuss the impact on DL and UL of loading AI/ML model to hardware before discussing whether and how to define requirements for Model update/transfer/delivery.** |
| [**R4-2316597**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316597.zip) | Samsung | *Requirement for data collection:***Proposal 1: For data collection for training purpose:**  **- Latency requirement shall not be further studied in RAN4;** **- The importance of defining accuracy requirement is use case-specific: Until a clear data collection procedure is defined in RAN1/2 (, i.e., probably in work item stage), there is no necessity for RAN4 to further study on accuracy requirement for data collection for training purpose.** **Proposal 2: For data collection for inference purpose,**  **- RAN4 requirement is defined only if the required data comes from other entities and the procedure of data collection is agreed in RAN1/2.**  **- The necessity of latency and/or accuracy requirement should be discussed for the particular use case for normative work.** **Proposal 3: the data collection for monitoring purpose, both accuracy and latency requirement can be specified for normative work, similar to RAN4 L3-measurement delay/accuracy requirement.***Handling of generalization***Proposal 4: For AI/ML generalization performance, RAN4 requirements/tests shall ensure the performance is maintained under different scenarios and/or configurations.** **Proposal 5: RAN4 define requirements/tests for AI/ML generalization performance:**  **- Training dataset: depends on DUT vendor;** **- Test/Inference dataset: Constructed by mixed datasets from multiple scenarios/configurations including Scenario#A/Configuration#A and a different dataset than Scenario#A/Configuration#A;**  **🡪 Study the necessity/feasibility of test scenario/configuration dynamic change during tests per (sub-)use case for normative work.** *Requirements for LCM***Proposal 6: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case for normative work.**  **- No need further RAN4 study until clear monitoring metrics/methods specified in other WGs.***RAN4 performance testing goals***Proposal 7: For RAN4 performance testing goal, Option 3 (Option 1 or 2 depending on test) is preferred to guarantee LCM procedure and performance gain in different use cases, both of which shall be considered in RAN4 in case-by-case manner.**  |
| [**R4-2316619**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316619.zip) | Nokia, Nokia Shanghai Bell | **On general aspects and testing goals:**1. A ML model can be identified, as agreed in RAN1, but only as an enabler of the feature/functionality and should not be tested in isolation from the feature/functionality.
2. RAN4 requirements and test procedures should be defined on the level of ML-enabled Functionality/Feature , i.e., model- specific requirements and tests shall be precluded.
3. The testing goal is to verify ~~whether a~~ the minimum performance of specific AI/ML ~~model~~ Functionality/feature ~~can be conducted in a proper way~~.
	* **~~FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)~~**
	* **FFS how to define ~~that~~ the minimum performance target(s) ~~model is properly conducted~~ (e.g., by defining AI/ML dedicated performance/core requirements associated with ~~model outputs~~ use cases)**
	* **FFS how to define ~~a static~~ the test scenarios/configurations (e.g., by defining a related testing dataset based on channel models in TR 38.901)**
	* **FFS whether to define non-static specific scenarios/configurations**
4. Definition of RAN4 requirements does not necessitate the definition of a single reference ML model for each ML-enabled Feature. Additional assumptions on the test parameters and/or functionality to align better the performance results can be discussed in per-use-case manner.
5. RAN4 to consider overhead when formulating performance requirements and comparing with legacy performance (Option 1).

**On data collection:**1. Clear distinction should be made in RAN4, when discussing the topic of data collection, namely data collection for training, inference, and monitoring purposes.
2. RAN4 to consider latency performance requirement of data collection only for ML inference and monitoring purposes. The data collection for functionality performance monitoring purposes can be further studied through LCM requirements.
3. RAN4 should study whether legacy measurement accuracy/error requirements are tolerable as the source of training data collection.

**On generalization/Robustness/Stability**Generalization aspects of AI/ML solutions introduce new challenges for RAN4 testing. ML-enabled Functionalities/Features may demonstrate performance degradation when the UE experiences radio scenario different from the scenario(s) used for training of ML models supporting the functionality/feature.1. For generalization verification of the ML-enabled Functionality/Feature, RAN4 to use the same (or very similar) test setup introduced for the minimum performance requirements.
2. Reference/typical scenario can be specified for AI/ML use cases. The performance of ML-enabled Functionalities/Features may variate in the conditions different from the typical/reference. Therefore, certain level of performance variation/degradation can be allowed in the alternative scenarios.
3. RAN4 should define new requirements and tests to guarantee a minimum performance of ML-enabled Functionality/Feature (use cases) in various radio conditions and network configurations.
4. RAN4 needs to define reference scenario as well as other/alternative scenarios with corresponding requirements to evaluate the generalization capabilities of the AI/ML-enabled features/functionalities for each of the (sub) use-cases studied in Release 18.
5. Another generalization aspect for ML-enabled Functionalities/Features is the capability of the UE or NW to address dynamic and/or semi-static changes of the radio conditions and/or configuration parameters.
6. RAN4 must design test configuration for verification of the generalization capabilities of an ML-enabled Functionality/Feature, not only in a set of fixed testing points/conditions but also when the conditions are changing from one scenario to another during the test.

**On UE processing capability and model complexity**1. In the current RAN4 testing setup, although multiple features might be simultaneously active on the device, only one feature at a time is tested to ensure its compliance with the requirements. In other words, all features would be tested individually, but testing concurrent features (e.g., beam prediction, CSI compression, CSI prediction, positioning, etc.) is not explicitly within the current scope of the test requirements.
2. RAN4 should tests the mutual impact of several simultaneously supported and active ML-enabled Functionalities/Features to ensure the absence of performance degradation.

When the air-interface and communication solutions are enabled by ML algorithms, it is desirable that both entities involved in the communication link are at least partially aware of the potential compute bottlenecks at the other node, to mutually optimize the relevant functionalities. 1. RAN4 shall assess the testability of and requirements on UE processing capabilities based on the performance indicators defined in the other WGs.

**On LCM and functional test**1. Based on the RAN1 proposed conclusion that Functionality-based LCM is the common baseline, RAN4 can start to design requirements and functional tests for Functionality-based LCM.
2. The requirements or functional tests for model-ID based LCM can be addressed by RAN4 in addition to the Functionality-based LCM requirements, after these are clarified.

In the ongoing RAN2 discussions there is no agreement so far on a ML Functional Framework for air-interface purposes, which RAN4 can use to design requirements and functional tests.1. For the use cases with UE-side models and UE-part of two-sided models, RAN4 to start the design of Functionality-based LCM related requirements and functional tests with the requirements needed for a Functionality operating in (ML) inference mode.
2. The ML inference and data collection functions located in the same node NW/gNB or UE, communicate with each other using implementation specific solutions i.e., do not require any RAN4 inter-operability testing.
3. The communication interfaces between the Functionality management (located at the NW-side) and the Functionality UE-side requires 3GPP specification to ensure inter-operability and testing/validation, for both the UE-side ML inference and UE-side monitoring data collection functions.
4. For UE-side models and UE-part of two-sided use cases, RAN4 needs to address the requirements and tests for the communication interface used for Functionality Management actions (activation/switching/fallback) towards the UE-side.
5. For UE-side models and UE-part of two-sided use cases, RAN4 to address the requirements and tests for the communication interface used for functionality monitoring data transfer between the data collection function at the UE-side and the data collection function at the gNB-side.
6. RAN4 to start the design of new LCM related requirements and functional tests starting with the transmission/exchange of Functionality monitoring data and Functionality management actions after these UE-gNB interfaces are agreed in RAN2.

**On post-deployment testing**1. Initial conformance testing of the AI/ML functionality cannot ensure the same level of performance for the devices in the field.
2. RAN4 should study a framework to enable post deployment tests for model updates and/or drift validation (and possible other use cases) and discuss at least the following non-mutually exclusive options:
	1. The changes/updates to the ML-enabled Functionality/Feature are tested and declared by the device vendor against RAN4 requirements before any deployment to the UE is performed.
	2. After deployment to the UE and before changed/updated ML-enabled Functionality/Feature is activated in the UE, a post-deployment validation is performed, e.g., a sanity check test loop is run, e.g., using the functionality performance monitoring and LCM activation/deactivation/switching procedures,
	3. At least one fallback/default Functionality/Feature that passed conformance testing must always be present in the device.
 |
| [**R4-2316856**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316856.zip) | Keysight Technologies UK Ltd | Feedback provided for the TR:1. In section 7.4.1 it is indicated that *“[requirements/tests for training will not be studied unless there is definition of training procedure.]”*. In our opinion, training data might be required to be discussed if RAN4 ends up fully defining test decoder. It doesn’t imply that requirements or test for training are required but specific discussions will be required to agree on training data even if training procedure is not defined.
2. In section 7.4.1 it is indicated that *“It should be considered for all tests (including LCM test) even not directly enforceable.”*. It is not clear what “*even not directly enforceable*” means. It is proposed to delete it.
3. Under *LCM related requirements/test* clause, there is a typo that should be corrected: “*The legacy framework for RRC/MAC-CE/DCI based core requirements*”.
4. Under *Training dataset* clause, it is indicated that “*Training dataset to be used for the device model training is left to implementation”.* We think thatsquare brackets should be added to this sentence until dual-sided test decoder option is chosen: if options 3 or 4 are finally used, RAN4 might need to define training dataset.
5. Under *Reference block diagram for 1-sided model* clause, should RAN4 references be updated/removed?
6. Under *Reference block diagram for 2-sided model* clause,
	1. should reference encoder/decoder be replaced by test encoder/decoder?
	2. Should we remove options 5 and 6 already down-selected?
 |

## Open issues summary

The open issues were grouped in the following sub-topics for further discussion:

1. Generalization - Model feasibility/robustness to different inputs
2. Generalization - Dynamically changing scenarios
3. Generalization
4. Testing goals
5. Latency requirements
6. TR updates
7. Updated to TR and LS to RAN1
8. UE Processing and feature concurrency
9. Tests for model updates (post-deployment)/drift validation

### Sub-topic 1-1

Generalization – high level

**Issue 1-1: Handling of Generalization**

* Proposals
	+ Option 1: Discuss generalization for each use case separately based on progress in other WGs, mainly RAN1
	+ Option 2: Discuss a general framework in RAN4 to establish generalization goals and candidate methods for handling it
* Recommended WF
	+ To be discussed

Mediatek: What is the difference between options. For option 1 we understand it means waiting for other WG.

Huawei: We have already had agreement since the first RAN4 meeting. RAN4 will study the feasibility and how to generalization based on RAN1 agreement. So far there is no generalization related agreement in TR. We prefer to Option 1 but it is duplication.

Nokia: We firstly should treat option 2 and then option 1. This is applicable to all the cases. Then we need discuss the parameters for each case.

Ericsson: We need both option 1 and option 2. In the end we need other WG input. But it does not mean we cannot do anything now.

Samsung: We see the value of Option 2. For RAN1 evaluation, they have several cases. The goal of generalization needs be clarified.

Vivo: For detailed aspects, we should study differently, e.g., the speed is important which will impact CSI.

OPPO: we can start with Option 2. Both are needed.

CMCC: We also see option 1 and option 2 are not contradict.

Qualcomm: How much value to discuss those options. For option 2, does anyone have concrete proposals?

ZTE: Option 1 is straightforward. RAN4 will define the requirements for generalization and we should base on RAN1 agreement.

### Sub-topic 1-2

Generalization goals for inference

**Issue 1-2: Generalization goals**

* Proposals
	+ Option 1: Verify whether the performance gain/minimum level of performance of AI/ML functionality/model can be achieved/maintain under various scenarios and/or configurations.
	+ Option 2: Verify whether the performance gain/minimum level of performance of AI/ML functionality/model can be achieved/maintain under certain scenarios and/or configurations (for which the model was designed) while not showing degradation in other scenarios and/or configurations
* Recommended WF
	+ To be discussed

Samsung: We use the different wording but we agree with Option 2. Comment on “other scenarios and/or configurations”.

Ericsson: Option 1 says that the same performance are expected. Option 2 says that there would be difference of performance on other scenarios. We do not need achieve the same performance in all the scenarios. Option 2 makes sense but it depends on the detailed scenario,

OPPO: If we only use the single model, the difference between option 1 and 2 is how many test cases we should consider. If we use the single model, then we can wait for WI phase. If we use multiple modes, then that is different from option 1 and 2.

Nokia: in general our view is that we do not need support the same level performance for all the cases. Option 1 is preferable. We should focus on minimum level of performance. For model functionality, can we really distinguish whether one or multiple modes will be used?

Apple: Option 2 is more realistic. However, how many scenarios should we consider? We should cover different typical scenarios. We can further identify the scenarios.

Qualcomm: Option 2 has more details compared to Option 1. If we think in more details, if we change the propagation condition for Doppler spread, do you expect the same SNR? How can we decide whether there is degradation or better performance? There are more details in Option 2 which need identification on how to test. Option 2 is starting point and more details need discussions.

Mediatek: Option 2 is more realistic. Option 1 has some values. For option 2, we need some level of performance for different sceanrios. We also agree with Apple how many scenarios we should consider. Agree with Qualcomm that the issue is complicated.

Vivo: All the test scenarios and configurations should be specified. Regarding how many, we can choose one or one set of scenarios and discuss further how many for each test.

Nokia: We prefer Option 2. Regarding Option 2, our concern is what is the degradation or same level performance.

Huawei: from our understanding, AI/ML is designed for the specific scenario. For functionality perspective, other WG has agreement that UE will report the AI functionality or model. If UE can support a certain scenario, we can select the scenarios accordingly.

ZTE: Compared to two options, shall we clarify the scenario or configurations?

**Agreement:**

* Verify whether the performance gain/minimum level of performance of AI/ML functionality/model can be achieved/maintain under the identified scenarios and/or configurations, while the performance won’t be significantly degraded in other scenarios and/or configurations
	+ FFS on details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality
	+ FFS on what the minimum level performance for each identified scenario and/or configuration is
	+ FFS on what the significant degradation for other scenarios and/or configurations is

### Sub-topic 1-3

Generalization handling - inference

**Issue 1-3: Handling of generalization in tests**

* Proposals
	+ Option 1: Signaling based LCM procedures and performance monitoring are considered in dedicated test cases and are excluded in tests verifying generalization and robustness. RAN4 defines multiple tests with different propagation conditions. In each of the test, TE configures the same ~~model ID~~ UE configuration, and therefore the same UE configuration  ~~AI/ML model~~ is tested under different propagation conditions to verify it’s generalizability and robustness. (environment differs in each test but not changing dynamically during the test)
	+ Option 2: RAN4 defines one test and changing different propagation conditions within the test. Therefore the same AI/ML model is tested under different propagation conditions to verify it’s generalizability and robustness. (environment changes during the test)
		- “morphing test concept”
		- Inference data set varied or picked from dataset used for another scenario
	+ Option 3: Use “Fuzzy side condition test” – side conditions vary during the test
	+ Option 4: Combinations of Option 1 , 2 and 3
* Recommended WF
	+ To be discussed

NTT DOCOMO: Prefer option 2. Option 1 and Option 2 have issues. For Option 1 how to define the test. Option 2, how to decide the range. If option 2 is not agreed, we can use Option3, which changes the conditions during test.

Mediatek: We do not prefer to change AI model during the test.

Qualcomm: Option 1. Mediatek suggestion is good. We want to keep the scope to test one model whether it is robust or not. What the difference between that UE uses different models and the conditions of test is changing. For Option 2. there are question how many tests we should have. Option 2 should be discussed in WI phase.

Ericsson: The wording for optioin 1~3 are model based testing. We should consider the functionality based testing. Option 1 is interesting to test the different condition. Option 2 is also interesting since it tests the model changing and test condition changing.

Vivo: We should check the static testing like Option 1. Compared to modelling changing in the different scenario, static testing is more practical.

Samsung: Similar view as Ericsson. Option 1 and Option 2 is like model based testing. We do not see how those options handle functionality based test.

Apple: Option 2 is the combinations of different option 1. The condition is kept changing during option 2. The only issue is the feasibility of the test for TE vendors. Test time and cost are our concern.

Nokia: Similar view as Ericsson. Both option 1 and 2 are interesting. It should be functionality based. In Option 1, it test the generalization aspects.

Huawei: Share the similar view as Mediatek and Qualcomm. For option 2, if the condition is varying the core requirement test will be involved. Option 1 is better.

OPPO: Option 1 and Option 3 are similar. The difference is that the model will be test in one or multiple. We prefer Option 2.

Intel: We see the fundamental difference of single fixed condition test or dynamic condition test. We prefer to the fixed test condition. We do not need rush for conclusion.

Qualcomm: we are proponent of Option 1.

Agreement:

* Take the modified Option 1 as the baseline
	+ Modified Option 1: Signaling based LCM procedures and performance monitoring are considered in dedicated test cases and are excluded in tests verifying generalization. RAN4 may define multiple tests with different conditions. In each of the test, TE configures the same specified UE configuration, and therefore the same specified UE configuration is tested under different conditions to verify it’s generalizability. (environment differs in each test but not changing dynamically during the test)
		- Specified UE configuration includes functionality and/or model ID if defined.
* FFS on Option 2
	+ In Option 2, change the same model ID to “the same specified UE configuration, which includes functionality and/or model ID if defined

### Sub-topic 1-4

*Testing goals*

**Issue 1-4: Testing goals**

* Proposals
	+ Option 1: The testing goal is to verify whether a specific AI/ML model can be conducted in a proper way.
		- FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)
		- FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)
	+ Option 2: The testing goal is to verify whether the performance gain of AI/ML model can be achieved for a static scenario/configuration.
		- FFS how to define a static scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)
		- FFS whether to define non-static specific scenarios/configurations
	+ Option 2a: The testing goal is to verify whether the minimum performance/performance gain of AI/ML ~~model/~~functionality/feature can be achieved for a static or non-static(dynamic) scenario/configuration.
		- FFS how to define the static test scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)
		- FFS how to define the minimum performance target(s) (e.g., by defining AI/ML dedicated performance/core requirements associated with use cases)
		- FFS how to define the non-static specific scenarios/configurations
	+ Option 3: Option 1 and Option 2/2a depending on the test
	+ Option 4: others, please provide some concrete proposals
* Recommended WF

Option 3

### Sub-topic 1-5

*Latency requirements*

**Issue 1-5: Latency requirements**

* Proposals
	+ Option 1: Latency requirements of data collection for model inference and monitoring should be considered and discussed per use case, subject to the output from RAN1/2.
	+ Option 2: RAN4 should study latency requirements for data collection of model monitoring, at least for positioning and CSI compression use cases.
	+ Option 3: RAN4 shall define the latency requirements based on RAN2’s agreements and the MAX total latency requirements can be: ****
	+ Option 4: Do not study latency requirements for training data collection, discuss latency requirements for any particular use case during WI as needed
	+ Option 5: Consider data collection latency requirements only for inference and monitoring
	+ Option 6: Other, please provide proposals
* Recommended WF

To be discussed

### Sub-topic 1-6

*TR updates*

**Issue 1-6: TR Updates**

* Proposals
	+ Option 1: TR update in R4-2315351
	+ Option 2: Keysight proposals in R4-2316856
* Recommended WF

Please provide comments to the proposed updates

### Sub-topic 1-7

*Update to TR and LS to RAN1*

**Issue 1-7: Update to TR and LS to RAN1**

* Proposals
	+ Option 1: Send the approved TP to RAN1 in the LS as attachment requesting RAN1 to incorporate the TP in TR38.843.
		- Addition agreed in previous meeting: Test encoder/decoder for TE: AI/ML model for UE encoder/gNB decoder implemented by TE.
* Recommended WF

To be discussed, is an LS needed?

### Sub-topic 1-8

*UE Processing and feature concurency*

**Issue 1-8: UE Processing and Feature Concurency**

* Proposals
	+ Option 1: RAN4 should tests the mutual impact of several simultaneously supported and active ML-enabled Functionalities/Features to ensure the absence of performance degradation.
	+ Option 2: Do not consider concurrent feature because there can be too many combinations
	+ Option 3: Leave this discussion to a work item phase when the features are known
	+ Option 4: Others
* Recommended WF

Option 3

### Sub-topic 1-9

*Tests for model updates (post-deployment)/drift validation*

**Issue 1-9: Post deployment testing**

* Proposals
	+ Option 1: RAN4 should study a framework to enable post deployment tests for model updates and/or drift validation(and possible other use cases)
		- Following options can be taken for reference in further discussion:
			* Option 1- a: The changes/updates to the ML-enabled Functionality/Feature are tested and declared by the device vendor against RAN4 requirements before any deployment to the UE is performed.
			* Option 1- b: After deployment to the UE and before changed/updated ML-enabled Functionality/Feature is activated in the UE, a post-deployment validation is performed, e.g., a sanity check test loop is run, e.g., using the functionality performance monitoring and LCM activation/deactivation/switching procedures,
			* Option 1- c: At least one fallback/default Functionality/Feature that passed conformance testing must always be present in the device.
			* Other options can also be discussed
	+ Option 2: RAN4 does not need to study such framework
	+ Option 3: others, please provide some proposals
* Recommended WF

Option 2

### Sub-topic 1-10

*“Ground truth” in RAN4*

*Ground truth is used in RAN1 simulations for comparisons/benchmarking, it would be useful to establish what is “ground truth” from a RAN4 point of view and whether this can be usable/accessible(used for comparisons) in RAN4 requirements or tests*

**Issue 1-10: “Ground truth” in RAN4**

* Proposals
	+ Option 1: Ground truth is the UE measured “raw data” at the baseband – channel estimation output, RSRP measurement output, etc
		- This is observed after part of the UE Rx processing chain
	+ Option 2: Ground truth is the input at the UE antenna ports – instantaneous channel at the UE antenna ports, instantaneous RSRP at the antenna port, etc
	+ Option 3: Discuss on a use case by use case basis
	+ Option 4: Others
* Recommended WF

Vivo: “Ground truth”, the input of AI model may be “ground truth” and we compare the input and output. We cannot get the ground truth for channel model.

Samsung: We need discuss what the purpose is. If it is for training, we do not we need discuss it. Related to use case, for some of use case, I do not think ground truth is needed.

Mediatek: It is related to RAN1. Ground truth would be useful for test metric. Question on the usefulness of ground truth. RAN4 just discuss the test and rely on measurements.

OPPO: Agree with Samsung. Ground truth depends on different cases.

Xiaomi: we need discuss. In RAN4 we only have genie or measured values. For AI, we have predicted value. There would be two step of comparison. Should we define the accuracy of measured compared to genie or of the predicted against measured values.

Apple: should we consider ground truth.

Moderator: it depends on the test metric.

# Topic #2: Specific Issues Related to Use Cases For AI/ML

This section contains the sub-topics regarding specific issues for the different use cases under study.

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2315104**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315104.zip) | CATT | **Proposal 1: Use throughput as eventual metric/KPI, and intermediate metrics/KPIs (SGCS, NMSE, etc.) should also be considered for CSI feedback.****Proposal 2: RAN4 waits for RAN1 conclusions on monitoring procedure and exact intermediate metrics/KPIs.** **Proposal 3: Intermediate KPIs (SCGS, NMSE, etc.) are feasible for CSI requirements or LCM. Pre-defined threshold may be required. The details need discussion case by case.** **Observation 1: The metrics/KPIs may be different for direct AI/ML positioning and AI/ML assisted positioning since the entity calculating UE positioning and the entity where the AI/ML models for positioning are deployed may be different for AI/ML assisted positioning case.****Proposal 4: Direct positioning accuracy can be used as metrics/KPIs for direct AI/ML positioning and legacy measurement quantities, e.g., ToA/AOA/LOS, for AI/ML assisted positioning. Exact measurement quantities can be FFS.****Proposal 5: CIR and PDP should not be the metrics/KPIs for positioning since they could be the inputs of AI/ML models for positioning and not easy to evaluate their accuracy.** **Proposal 6: RAN4 not to define requirements for model delivery/update/transfer unless it is requested to do so.** |
| [**R4-2315323**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315323.zip) | CMCC | ***Proposal 1: for CSI compression, except throughput, cosine similarity can be considered as KPI/test metrics for RAN4 study.******Proposal 2: for beam management, it is proposed to consider both RSRP accuracy and beam prediction accuracy as KPIs.******Proposal 3: for direct AI/ML positioning, it is proposed to consider positioning accuracy: Ground truth vs. reported as metric.******Proposal 4: it is proposed to define requirements for model monitoring in LCM. And it is proposed to study whether to use same or different metric as that for inference per use case.*** |
| [**R4-2315408**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315408.zip) | xiaomi | **Observation 1: Many metrics for different motivation for beam management are defined in RAN1.****Proposal 1: RAN4 to discuss test metric for different motivation for beam management separately.****Proposal 2: RAN4 needs to define test metric for both beam prediction accuracy and RSRP prediction accuracy.****Observation 2: If ideal L1-RSRP refers to measured L1-RSRP, RSRP delta defined in RAN1 is relative RSRP accuracy which is compared with measured RSRP, it can’t reflect the actual RSRP predication accuracy(absolute RSRP accuracy).****Proposal 3: For RSRP prediction accuracy metric, RAN4 to discuss the definition of ideal L1-RSRP specified in RAN1. If ideal L1-RSRP refers to measured L1-RSRP, RAN4 needs to define absolute RSRP accuracy, i.e. between predicted RSRP with Genie RSRP for the same beam.****Proposal 4: For Beam prediction accuracy metric, RAN4 needs to discuss whether to test both beam index based and RSRP based metric or only choose one type to test.****Observation 3: Beam prediction accuracy can be evaluated in terms of RSRP difference or beam index difference.** **Observation 4: For both types of test metric, Genie-aided beam will be used as benchmark.****Proposal 5: For Beam prediction accuracy metric, RAN4 needs to discuss how to design Genie-aided beam, whether UE or Test equipment provide the Genie-aided Top-1/K beam.****Observation 5: For Beam prediction accuracy metric, if measured L1-RSRP is used as ideal L1-RSRP, then the RSRP delta defined in RAN1 will include actual RSRP delta between two beams plus measurement error.****Proposal 6: For Beam prediction accuracy metric, RAN4 to discuss the definition of ideal L1-RSRP specified in RAN1 and how to design testcase.** |
| [**R4-2315481**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315481.zip) | Apple | 1. *SGCS, NMSE need data from UE processing chain – Eigen vectors, channel estimates, etc which are not typical output.*
2. *It is not clear how intermediate KPIs like SGCS, NMSE can used as metric in performance testing.*
3. **Deprioritize using intermediate KPIs – SGCS and NMSE as test metrics for RAN4 requirements since it is not practical and feasible.**
4. **Capture the confirmed option(s) for test metric for CSI requirements in the TR.**
5. *For beam prediction it is not feasible to choose one metric for requirements at this stage.*
6. **Capture all the feasible options for test metric for beam prediction requirements.**
 |
| [**R4-2315601**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315601.zip) | Ericsson | [Observation 1 The approach used for deciding the test decoder is linked to the discussions on training and interoperability.](#_Toc146733296)[Observation 2 The need to ensure interoperability between different encoders and decoders of different vendors, including testing, represents a significant complexity for the industry and a step away from standardization.](#_Toc146733297)[Observation 3 It may be that the overall performance benefit obtainable from being able to operate with different decoder/encoders does not justify the complexity of supporting interoperability.](#_Toc146733298)[Observation 4 The work involved in standardizing a decoder, whilst significant, may still not be as large as the cost to the industry of supporting large numbers of encoder/decoder combinations and interoperability.](#_Toc146733299)[Observation 5 A standardized decoder would need to provide comparable performance for operation after compiling across multiple different platforms.](#_Toc146733300)[Observation 6 Standardization of a decoder is quite different to today, since the models are not human readable and debateable. Discussion would presumably focus on how transferable the models would be between different platforms.](#_Toc146733301)[Observation 7 The most convincing case for not standardizing a decoder is where the gNB may trade off complexity and performance in different ways. In that case, since it is the gNB trading cost and performance it makes sense for test decoders and potentially API for reference encoder to be supplied from the network side.](#_Toc146733302)[Observation 8 Even if there is a complexity/cost trade-off, standardization of several test decoders with different complexity levels could be considered.](#_Toc146733303)[Observation 9 The conclusions of the SI on interoperability/testing for CSI compression may not be generally applicable for 2-sided models, since it depends on the variation of performance of encoders / decoders.](#_Toc146733304)[Observation 10 It is assumed when filling in the table that the test decoder/encoder will in practice need to be closely matched to the actually used decoder / encoder, and hence the selection of encoder/decoder impacts training and collaboration options.](#_Toc146733305)[Observation 11 Care should be taken when reading Option 1 and Option 2 for BS testing in two-sided CSI compression. Assumptions should be clearly stated to ensure alignment and decrease confusion.](#_Toc146733306)[Observation 12 RAN4 may need to introduce performance metric calculation or reporting requirements based on the RAN1 design in WI.](#_Toc146733307)[Observation 13 RAN4 may need to define type 2 performance monitoring requirements if Ran1 agrees to adapt that method in WI phase.](#_Toc146733308)[Observation 14 If the data collection requires UE need to perform measurements other than legacy L1-RSRP measurements, RAN4 need to define the measurement period requirements for the data collection.](#_Toc146733309)Based on the discussion in the previous sections we propose the following:[Proposal 1 RAN4 to define LCM requirements in the WI phase.](#_Toc146733371)[Proposal 2 Consider the information in sections 2.2.1. and 2.2.2 as responses to the table contents.](#_Toc146733372)[Proposal 3 RAN4 to introduce performance monitoring requirements during the WI phase.](#_Toc146733373)[Proposal 4 RAN4 to study introducing requirements for beam pair(s) if RAN1 agreed to introduce it in WI phase.](#_Toc146733374)[Proposal 5 Ran4 to define requirements for confidence metric calculation and reporting to make sure same baseline performance for all the UE.](#_Toc146733375)[Proposal 6 RAN4 to study improvement to L1-RSRP measurement accuracy and the conditions under which it can be improved for model input for better model training/inference.](#_Toc146733376)[Proposal 7 RAN4 to study requirements in estimating RSRP accuracy and beam prediction accuracy given RSRP measurement inaccuracies](#_Toc146733377)[Proposal 8 RAN4 to not consider *positioning accuracy: ground truth vs. reported* as one of the positioning KPIs/metrics for positioning use case.](#_Toc146733378)[Proposal 9 Do not consider the accuracy of path phase as one of the positioning KPIs/metrics for AI/ML based positioning.](#_Toc146733379)[Proposal 10 RAN4 to first evaluate feasibility aspect before considering the accuracy of LoS/NLoS indication as one of the KPIs/metrics for AI/ML based positioning.](#_Toc146733380)[Proposal 11 Accuracy of RSTD as output of AI/ML model, at least for positioning sub-use case 2a, shall be considered as one of the KPIs/metrics for positioning use case in RAN4 discussion.](#_Toc146733381)[Proposal 12 RAN4 to not consider PRS-RSRP accuracy as one of the KPIs/metrics for positioning use case in RAN4 discussion, unless PRS-RSRP is one of the potential outputs of AI/ML model used for positioning.](#_Toc146733382)[Proposal 13 RAN4 to discuss other positioning KPIs/metrics, if found relevant, for AI/ML based positioning during the WI phase.](#_Toc146733383)[Proposal 14 Accuracy requirement for label data (corresponding to model output) needs to be defined if collection of training data over the air interface is agreed to be standardized.](#_Toc146733384)[Proposal 15 Accuracy requirement for measurement data (corresponding to model input) needs to be defined.](#_Toc146733385)[Proposal 16 For model inference, reporting format (quantization range and granularity) of measurement data need to be defined by RAN4 for positioning sub-use case 2b (LPP interface) and sub-use case 3b (NRPPa interface). Details to be further discussed during the WI phase.](#_Toc146733386) |
| [**R4-2315727**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315727.zip) | vivo | ***Proposal 1: RAN4 to use throughput be the metrics for CSI requirements/tests for model inference performance testing and further study the intermediate KPIs in the future release if necessary.******Proposal 2: RAN4 to use RSRP accuracy as the baseline KPI to evaluate AI/ML based beam management inference performance******Proposal 3: No positioning accuracy requirements for direct AI/ML positioning are defined.******Proposal 4: RAN4 is to study whether requirements/tests should be defined for potential new measurements for channel estimation, including the model input CIR/PDP and existing measurements used for direct AI/ML positioning.******Proposal 5: RAN4 to study potential requirements for new and existing measurements for AI/ML assisted positioning, at least for ToA, RSTD and RSRPP.******Proposal 6: For AI/ML assisted positioning using existing measurements, e.g., RSTD, RSRPP, legacy requirements could be used as starting point.******Proposal 7: Requirements and tests for model transfer/delivery is necessary to be specified and more progress from other WGs is needed.******Proposal 8: RAN4 to study how to define delay requirement and accuracy requirement for model monitoring for different monitoring procedure, e.g., monitoring decision at UE side or gNB side.******Proposal 9: Model monitoring requirements should be discussed per use case basis.*** |
| [**R4-2315995**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315995.zip) | Huawei,HiSilicon | ***Proposal 1:*** The down-selection of RAN4 identified potential test metric in AI CSI compression needs to wait RAN1/RAN2 progress on related signaling/procedure definition related to model transfer/delivery. ***Proposal 2:*** If the model under test of the DUT is transferred over air interface signaling from the opposite site, then the throughput may not be applicable. In this case, intermediate KPIs seems to be more applicable. However, how to obtain the expected model output is still an open issue.***Proposal 3:*** If the model transfer/delivery of the model under test to the DUT is spec transparent, then the throughput may be applicable. However, how to eliminate the effect of the operations from the opposite side is still an open issue.***Proposal 4:*** RAN4 studies the potential KPIs/test metrics in AI/ML spatial-frequency CSI compression in Table 2.1.

|  |
| --- |
| Table 2.1 KPIs/test metrics in AI/ML Spatial-frequency CSI compression  |
| Test Objective | Type 1 NW Joint Training | Type 3 Separate Training |
| gNB | UE | gNB | UE |
| KPIs/Test Metrics | * Throughput
* Accuracy of CSI decompression (intermediate KPIs, e.g. cosine similarity)
 | Accuracy of CSI compression (intermediate KPIs, e.g. cosine similarity) | * Throughput
* Accuracy of CSI decompression (intermediate KPIs, e.g. cosine similarity)
 | * Throughput
* Accuracy of CSI compression (intermediate KPIs, e.g. cosine similarity)
 |

***Proposal 5:*** If throughput is the test metric for AI CSI prediction, how to ensure that the testing dataset aligns well with training dataset is still an open issue.***Proposal 6:*** If intermediate KPI is the test metric for AI CSI prediction, which entity provides the ideal CSI is still an open issue.***Proposal 7:*** RAN4 studies the potential KPIs/test metrics in AI/ML Temporal CSI prediction in Table 2.1.2.

|  |
| --- |
| Table 2.2 KPIs/test metrics in AI/ML Temporal CSI prediction  |
| Test Objective | AI/ML model @ UE |
| gNB | UE |
| KPIs/Test Metrics | / | * Throughput
* Accuracy of CSI prediction (intermediate KPIs, e.g. cosine similarity)
 |

 |
| [**R4-2316229**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316229.zip) | OPPO | **Proposal 1: For RAN4 performance test, at least two aspects need to be considered:** * **Model/functionality input, the focus of testing is on whether the input information of the model/functionality could be correctly obtained**
* **Model/functionality output, the focus of testing is on whether the performance of a given model/functionality could be guaranteed**

**Proposal 2: RAN4 does not need to study requirements/tests for AI/ML model delivery in each use case.** **Proposal 3: RAN4 does not need to study requirements/tests for AI/ML model update, AI/ML model transfer in each use case.**  **Note: if other WG defines the model update procedure or model transfer procedure, RAN4 may need to study corresponding requirements for it.****Proposal 4: For AI/ML based CSI feedback, performance requirement on CSI model/functionality input (e.g. CSI-RS measurement accuracy) should be studied.****Proposal 5: For CSI inference performance,** **throughput should be used to evaluate the model inference performance, and existed RAN4 test examples for “reporting of PMI” can be reused or serve as a reference. Requirement of γ and test settings can be reused or updated.****Proposal 6: For CSI performance monitoring, UE can monitor and estimate the performance of AI/ML based CSI model/functionality through Hypothetical BLER or intermediate KPIs, e.g. SGCS.**  **- FFS how to perform cell level AI/ML model/functionality performance monitoring****Proposal 7: RAN4 R18 does not need to study requirements/tests for CSI model update/transfer/delivery****Proposal 8: For AI/ML based BM, performance requirement on BM model/functionality input (e.g. beam measurement accuracy) should be studied.****Proposal 9: For BM inference performance, RSRP accuracy of the predicted beam(s) should be used for BM model inference performance tests.****Proposal 10: For BM performance monitoring, UE can monitor and estimate the performance of AI/ML based CSI model/functionality through RSRP accuracy or other intermediate KPIs, e.g. Beam prediction accuracy.** **- FFS how to perform cell level AI/ML model/functionality performance monitoring****Proposal 11: RAN4 R18 does not need to study requirements/tests for BM model update/transfer/delivery****Proposal 12: For AI/ML based Positioning, performance requirement on Positioning model/functionality input (e.g. measurement accuracy of CIR/PDP/RSRP /RSTD) should be studied for all AI/ML positioning cases, i.e., for case 1/2a/2b/3a/3b.****Proposal 13: For positioning case1, positioning accuracy should be utilized as the KPI to test the model/functionality output.****Proposal 14: Regarding RAN4 tests on positioning accuracy, using positioning test data set(s) which including model input-related data (e.g. measurement results of PRS that matched with positions, or model inputs obtained from the measurement results) and positioning label data, prepared in advance by TE could be considered.** **Proposal 15: For case2 and UE-side AI/ML assisted positioning in case1, considering the accuracy for intermediate results could be considered as the KPI to test the model/functionality output.** **Note: most of the intermediate measurement results agreed in RAN1 are not realistic measurement results. They are from AI/ML model output(with non-linear processing). How to get the label data(the expected intermediate measurement results) and test these intermediate measurement results is also a new and a more challenge issue in RAN4. Whether option 2 to option 6 agreed in RAN4#107 meeting can be used in RAN4 tests as a KPI metric should be further analysed. Feasibility of using these candidate metrics should be clarified.****Proposal 16: For positioning case2b,3a,3b (cases without UE-side model), not necessary to test the Positioning model/functionality outputs.** |
| [**R4-2316393**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316393.zip) | Nokia, Nokia Shanghai Bell | 1. The CSI use case impacts precoding matrix part of the CSI reporting requirements.
2. RAN4 should further study the impacts of AI/ML-enabled CSI use cases on the UE performance requirements in TS 38.101-4. A specific new target value of γ (gamma) for AI/ML-enabled CSI use cases can be envisaged.

A new relative throughput performance indicator can be introduced for AI/ML-enabled CSI use cases. 1. RAN4 should further study if a new relative throughput performance indicator would be more suitable for AI/ML-enabled CSI use case, other than the legacy γ (gamma).

For CSI prediction performance monitoring, RAN1 is already discussing about network side performance monitoring option (Type 2). The same approach can be applied to test the CSI prediction accuracy as a performance KPI at RAN4.CSI Prediction Accuracy as a KPI can be testable using the existing test interfaces with minimum change in the TE.1. Apart from throughput, CSI prediction accuracy should be used as an additional performance KPI for RAN4 requirements.
2. RAN4 to consider the difference between the predicted L1-RSRP of Top-K predicted beam and the ideal L1-RSRP of the same beam as metric/KPI for L1-RSRP accuracy requirements/tests.

For BM-Case1 and BM-Case2, Top-1(%) and Top-K/1(%) were agreed in RAN1 and baseline related KPIs for AI/ML performance evaluation in beam management use-case.1. RAN4 to consider Top-1(%) and Top-K/1(%) as metrics/KPIs for beam prediction accuracy requirements/tests.

For BM-Case1 DL Tx beam prediction, when Set B is the subset of Set A, the KPIs of the beam ID(s) prediction need to be verified. For BM-Case1 DL Tx beam prediction, when Set B is the subset of Set A, further study needs to be done for the selection of target values of prediction accuracy KPIs.1. RAN4 should further study the test mechanism for AI/ML based Top-1(%), Top-K/1(%) beam IDs prediction for BM-Case1 DL Tx beam prediction, when Set B is subset of Set A.
2. RAN4 should further study the test mechanism for AI/ML based Top-1(%), Top-K/1(%) beam IDs prediction for BM-Case1 DL Tx beam prediction. Further study should be done for the selection of target values of prediction accuracy KPIs, when Set B is subset of Set A.

Positioning coordinates are inference output of AI/ML model functionality in case of UE based direct AI/ML Positioning.1. Positioning accuracy should be considered as Test metric/KPI in RAN4 for inference validation for UE based direct AI/ML Positioning.

Positioning accuracy can be verified based on the ground truth which may consist of the location points with known positioning co-ordinates (e.g., PRU or GNSS based).1. Some new test mechanism should be studied in RAN4 for validation of Positioning accuracy KPI based on the ground truth which would consist of known positioning co-ordinates.

For Assisted AIML positioning, intermediate KPIs such as LOS/NLOS have direct impact on the positioning accuracy.1. For Assisted AIML Positioning, the intermediate KPIs (e.g., LOS/NLOS) need to be considered for validating the positioning accuracy.

For Assisted AI/ML Positioning, PRS RSRP/ RSRPP, UE Rx-Tx or RSTD inference from AI/ML model/ functionality serving as an input for the positioning algorithm at UE/ LMF will have impact on the positioning accuracy. Hence, PRS RSRP/ RSRPP, UE Rx-Tx and RSTD should be considered as intermediate KPIs/ features and the existing measurement accuracy requirements for these KPIs should be further analyzed and adapted for AI/ML based model/ functionality.1. For Assisted AI/ML Positioning, the existing measurement accuracy requirements of the UE Rx-Tx and RSTD should be considered as option 2 and further analyzed and adapted for AI/ML based model/ functionality.
2. CIR/PDP should not be considered as the intermediate KPIs (option 3) since they can only be inputs to AI/ML functionality/model but cannot be an output. Hence it should be removed from option 3.
3. For Assisted AI/ML Positioning, the existing measurement accuracy requirements of the RSRP and RSRPP (newly added) should be considered as option 3 and further analyzed and adapted for AI/ML based model/ functionality.

For UE-sided model and UE sided monitoring, the UE needs to detect the performance degradation/improvement with respect to the predefined threshold for different KPIs in BM-Case 1 and BM-Case 2For UE-sided model and NW sided monitoring, the test needs to ensure that the UE performs with respect to the request from NW regarding LCM operations. 1. For BM-Case 1 and BM-Case 2, RAN4 should further study the test mechanism for the performance metric(s) of AI/ML functionality-based LCM at UE side and NW side.

Generalization aspects should be tested in RAN4 for BM use case. 1. Some reference radio conditions, and configuration/parameters settings should be identified for BM use case and different scenarios can be mutually agreed to be tested in addition of reference conditions and configurations.
2. For BM-Case1 and BM-Case2, RAN4 should study the impact of tolerance margin in legacy L1-RSRP measurement accuracy requirements on the performance of AI/ML based BM.
 |
| [**R4-2316468**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316468.zip) | MediaTek inc. | ***Observation 1***: **In RAN4#107, it was agreed to use throughput as the test metric to define CSI requirements.*****Observation 2***: **In RAN4#107, it was agreed to consider other metrics or intermediate KPIs if throughput is not applicable or significant disadvantage is observed.*****Observation 3***: **Other test metrics and intermediate KPIs are not precluded in the SI stage according to the agreements in RAN4#107.*****Proposal 1***: **In the stage of SI, RAN4 can capture possible test metrics, including throughput, intermediate KPIs or other test metrics in the TR. However, we suggest following the agreements in RAN4#107 to use throughput as the test metric to define CSI requirements in the WI stage. If throughput is not applicable or significant disadvantage is observed, RAN4 can consider other metrics or intermediate KPIs. RAN4 can then determine the test metric after RAN1’s work in Rel-19 WI.**

|  |  |  |
| --- | --- | --- |
| Option 1: Only use throughput (absolute or relative) | Option 2: Use throughput and other intermediate metrics/KPIs (SGCS, NMSE, etc) | Option 3: use throughput and overhead |
| * Can use ACK/NACK in legacy design to calculate throughput
 | * Intermediate KPI is only utilized for AI/ML model performance calibration in RAN1
* Need the ground truth which depends on physical layer design depends in RAN1
 | * It may be difficult to introduce the requirement by considering both throughput and overhead
* The overhead can be used as a side condition
 |

***Observation 4***: **It is possible to get the ideal L1-RSRP during the test in FR1.*****Proposal 2***: **Discuss whether to consider FR1 in AI/ML based beam prediction.*****Proposal 3***: **RSRP accuracy refers to the accuracy of predicted RSRP compared to measured RSRP in FR2 and ideal RSRP in FR1.*****Proposal 4***: **Different metrics can be used for different model types unless RAN1 makes down-selection on the model types.*****Proposal 5***: **Regarding the metric(s) for Beam prediction requirements/tests, we can keep all options and have further discussions in the WI based on the table below.**

|  |  |  |
| --- | --- | --- |
| Option 1: the accuracy of predicted RSRP compared to measured RSRP in FR2 and ideal RSRP in FR1 | Option 2: beam prediction accuracy: Top-1(%), Top-K(%) | Option 3: The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB |
| Not applicable to some implementation of AI model (e.g., not output of predicted L1-RSRP | Applicable to all studied AI models  | Applicable to all studied AI models |
| Feasible to define both accuracy requirements and test cases | Feasible to define test cases. No need to define accuracy requirements | Feasible to define both accuracy requirements and test cases |

 |
| [**R4-2316595**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316595.zip) | Samsung | **Observation 1: Throughput is considered as a promising test metric for both CSI compression and CSI prediction.****Observation 2: Only reason for RAN1 to have the intermediate KPIs is to simplify their evaluation without actual testing.****Proposal 1: RAN4 is recommended to focus only on the throughput rather than those intermediate KPIs for model inference discussion.****Observation 3: It has been shown that the CSI compression** **as the two-sided model can be much impacted by training assumption, and the cost that may necessitate offline multi-vendor collaboration for the limited gain.****Observation 4: Most options for the beam management inference performance would be good metrics/KPIs to have for further performance verification.****Proposal 2: RAN4 should study at least the network side model for DL Tx beam prediction for both spatial domain and temporal domain beam management cases without the procedure of model delivery/update/transfer.****Observation 5: For the direct AI/ML positioning, it would be too premature not to define or study the test metric in RAN4. RAN4 can have further discussions on alternative ways to evaluate this AI/ML scenario with legacy features or products for the reference, e.g., reference positioning entity.****Observation 6: LOS/NLOS indicator or PRS RSRP is preferred for the AI/ML assisted positioning as these metrics are commonly used in the RAN1 study.****Observation 7: Both direct AI/ML positioning and AI/ML assisted positioning have been shown that they can significantly improve the positioning accuracy compared to existing RAT-dependent positioning methods.****Proposal 3: Both sub use cases should be kept considering NW and UE side model without model transfer for further discussion.** |
| [**R4-2316801**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316801.zip) | Google Inc. | **Proposal 1: Defer the discussion for the requirements of AI/ML model monitoring until RAN1 has conclusion.****Proposal 2: The relative throughput should be used as the test metric in first priority to evaluate model inference performance for AI/ML CSI feedback enhancement use cases.** **Proposal 3: Since the intermediated KPIs are not stable when transmission rank >1 according to RAN1 simulation, it is proposed to deprioritize the intermediated KPIs to be used as the test metric to evaluate model inference performance for AI/ML CSI feedback enhancement use cases.****Proposal 4: Reference decoder(s) for the two-side model testing should be provided by the vendor of the encoder.** |

## Open issues summary

The open issues were grouped in the following sub-topics for further discussion:

1. Metrics/KPIs for CSI requirements/tests
2. Metrics/KPIs for Beam prediction requirements/tests
3. Metrics/KPIs for positioning requirements/tests
4. Requirements for model transfer/delivery/update
5. Accuracy requirements for measurement data or label data
6. Reporting format for positioning
7. Throughput applicability for transferred models

### Sub-topic 2-1

*Metrics/KPIs for CSI requirements/tests*

Different metrics/KPIs are being proposed for the CSI use case. While there seems to be consensus that throughput(relative throughput) should be used as one of the metrics, some companies are proposing to also further study some intermediate metrics(SGCS, NMSE, etc) others are proposing to discard such KPIs

**Issue 2-1: Metrics/KPIs for CSI requirements/tests**

Proposals

* + Option 1: Throughput/relative throughput
	+ Option 2: SGCS, NMSE
	+ Option 3: CSI prediction accuracy
* Recommended WF

To be discussed which options should be kept for further study. Companies proposing to use intermediate KPIs should provide more details on how these can be used as metric in practical RAN4 tests

Moderators: for SGCS, do we need ground truth or not?

Mediatek: we have already had agreement. Option 1 is baseline. We just follow that for completion of SI. In RAN1 they have also genie information for that. We need follow the previous agreement.

Apple: We can use option 1 as baseline. For relative we have taken the existing one as benchmark. For option 3, we can still find the genie. Agree with companies, we can first exclude option 2 at this stage.

Vivo: we have already had agreement. During study phase, the existing agreement is enough. For SGCS and NMSE, the ground truth is the raw data in the channel. RAN1 is discussing some mechanism to achieve the raw data.

Nokia: it is related to issue 1-10. The issue is complicated. It is possible to verify CSI compared to measured data. Ground truth needs be discussed based on case. We would like to keep option 1 and 3.

Moderator: Option 1 is baseline. Do we need keep studying Option 2 and option 3? it is strange to keep SGCS but have not feasible way to use it.

Ericsson: We are OK to do option 1 not to study option 2 and 3. Option 2/3 we may need study ground truth.

Intel: We are OK to go with Option 1 as baseline. OK to remove Option 2.

Agreement:

* For **Metrics/KPIs for CSI requirements/tests, use Option 1 as baseline**
	+ For Option 3, further discuss the feasibility to define the CSI prediction accuracy in the WI phase.
* FFS for monitoring metrics

### Sub-topic 2-2

*Metrics/KPIs for Beam prediction requirements/tests*

In RAN4#108 it was agreed to have the following candidate metrics/KPIs for beam prediction:

1. RSRP accuracy
2. beam prediction accuracy :Top-1(%), Top-K(%)
3. The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB,
	* + Related measurement accuracy can be considered to determine x
4. combinations of above options

It was also agreed that overhead/latency reduction can be considered as a side condition

**Issue 2-2: Metrics/KPIs for Beam prediction requirements/tests**

* Proposals
	+ Option 1: further downselect one/more of the above
	+ Option 2: document all the above in the TR as possible metrics
	+ Option 3: add other metrics?
* Recommended WF

Discuss the next steps for beam prediction KPIs

Xiaomi: For RSRP accuracy, it is related to beam measurement. Beam prediction is just one metric. There are two types: RSRP prediction accuracy; beam index difference. What is RSRP accuracy?

Qualcomm: Beam index and RSRP prediction are covered by the existing options.

Nokia: we are OK with Option 2.

Ericsson: keep options open. For option 2, can we keep studying options?

Apple: now it is moving target. RAN1 is still discussing it. If RSRP is supported, we can use the legacy requirement, i.e., just checking the RSRP.

Agreement: use option 2 as baseline to prepare TP.

### Sub-topic 2-3

*Metrics/KPIs for positioning requirements/tests*

There are several competing proposals on the figures of merit to be used for positioning

**Issue 2-3: Metrics/KPIs for positioning requirements/tests**

* Proposals
	+ Option 1: ground truth vs. reported location
	+ Option 2: CIR/PDP, channel estimation accuracy
	+ Option 3: ToA, RSTD and RSRP, and RSRPP
	+ Option 4: others (e.g., intermediate KPIs, LoS/NLoS)/combinations of the above
* Recommended WF

To be discussed – also needs to be clarified whether new requirements would be needed and how these can be defined

OPPO: we have six different cases for positioning. We also have different combinations. Different cases may need different metrics. Now we can have high level agreements. All metrics need be considered. For Option3, we have concern on the feasibility of it. The KPI is the output of AI model. It is hard to get the label data. How to define it.

Apple: for positioning, we should follow the legacy framework. We need wait for conclusion of RAN1.

Intel: we have different cases for position features. We need define the feasibility for each of them. Agree with Apple that we should take the legacy measurement as baseline. Regarding anything else which UE does not report, we should be careful for them. Location reporting we should be careful about.

Vivo: It is difficult to down-select. For option 1 it is difficult to do this. Different AI models are used for different deployment. For option 3, it depends on how to derive ground truth.

Nokia: in our view, Option 2 can be down-selected.

Samsung: in the end, all the six cases may not be defined. We can wait for RAN plenary decision.

Agreement: Prepare TP to capture the agreed options for metrics in the previous meetings

### Sub-topic 2-4

*Model delivery/update/transfer requirements*

Some documents are discussing whether there is a need to develop requirements for model deliver/update/transfer

**Issue 2-4: Model delivery/update/transfer requirements**

* Proposals
	+ Option 1: RAN4 to study requirements for model delivery/update/transfer
	+ Option 2: RAN4 does not need to study such requirements
	+ Option 3: study necessity of requirements only if corresponding procedures are defined by other WGs
* Recommended WF

To be discussed

### Sub-topic 2-5

*Accuracy requirements for measurement data or label data*

Some papers are discussing the possibility/necessity to define requirements for measurement data (e.g. measurements report by the UE) or labeled data

**Issue 2-5: Accuracy requirements for measurement data or label data**

* Proposals
	+ Option 1: RAN4 should study the possibility of defining requirements for measurement data or labelled data
	+ Option 2: No need to do anything else other than existing requirements
	+ Option 3: Other options
* Recommended WF
	+ To be discussed

Ericsson: It depends on the framework whether the measurement/collected data will be specified.

Qualcomm: Accuracy requirement for measured data or labelled data is important. Maybe legacy requirement may be applied. But if it is new, we need further discussion.

Ericsson: even for legacy requirement, we need discuss whether the tighter requirement is needed.

Nokia: is it the same issue for data collection for training. What is the difference? Agree with Ericsson.

Qualcomm: We do not restrict any.

Tentative agreement:

* If the data collection is specified, then the accuracy requirements
	+ For measurement data or label data based on legacy procedures, the legacy requirements can be used as baseline and enhancement can be discussed per use case basis in WI stage.
	+ For new reporting or measurement, the requirements can be discussed per use case basis in WI stage.

### Sub-topic 2-6

*Reporting format for positioning*

**Issue 2-6: Reporting format for positioning**

* Proposals
	+ Option 1: reporting format (quantization range and granularity) of measurement data need to be defined by RAN4 for positioning sub-use case 2b (LPP interface) and sub-use case 3b (NRPPa interface). Details to be further discussed during the WI phase.
	+ Option 2: Such definition is not needed in RAN4
	+ Option 3: leave such discussion to WI phase (if needed), no need to study feasibility
* Recommended WF
	+ To be discussed

### Sub-topic 2-7

*Throughput/KPIs applicability depending on model source*

**Issue 2-7: Throughput/KPI applicability depending on model source**

* Proposals
	+ Option 1: If the model is provided the “opposite side”(for example, provided by the network and transferred to the UE), then throughput test would not be applicable because performance has to be ensured by the model provider (i.e “opposite side”). There is still an open issue on how to ensure model integrity and that the performance is tested/met
	+ Option 2: If the model is provided by the device vendor ,(for example, provided by the UE vendor to the UE) then throughput or other KPIs are applicable since the device vendor should guarantee the performance. FFS how to guarantee the performance
	+ Option 3: Others
* Recommended WF
	+ To be discussed

# Topic #3: Interoperability and testability aspect

This section contains the sub-topics regarding interoperability and testability.

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2315066**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315066.zip) | CAICT | **Proposal 1: Following updates (shown in figure 2 and 3) to reference block diagram of one-sided model could be considered.**Figure 2 updates to reference block diagram of one-sided model (DUT-side monitoring)Figure 3 updates to reference block diagram of one-sided model (TE side monitoring)**Proposal 2: Following updates (shown in figure 4 and 5) to reference block diagram of two-sided model could be considered.**Figure 4 updates to reference block diagram of two-sided model (DUT-side monitoring with proxy model)Figure 5 updates to reference block diagram of two-side model (TE-side monitoring with ground-truth CSI reporting)**Observation 1: The details of option 4 and 6 for reference decoder need to be clarified to have a common understanding.****Proposal 3: Our views on different options of test decoder for 2-sided model are listed in the table 1.** *–***Table 1 Views on different testing options for two-sided model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder | UE vendor | gNB vendor |  Specification | UE/gNB/TE + specification |
| Source of decoder training data | Up to UE implement and alignment of training data between UE and TE might be required. | Up to gNB implementation and alignment of training data between UE and gNB might be required. | N/A | Up to implementation and alignment between UE, gNB and TE vendor might be required. |
| DUT vendor knowledge of the test decoder | Full knowledge | No knowledge or partial knowledge depend on the alignment between gNB/TE and UE vendor | Full knowledge | Partial knowledge  |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | Type 1/2/3, depending on UE implementation | Type 1/2/3, depending on offline alignment between UE and gNB/TE vendors. |  Type 1/2/3 | Type 1/2/3, depending on UE implementation or offline alignment between gNB and UE vendors. |
| Test decoder verification procedure at TE and/or DUT | TE needs to verify the decoder runs properly. | UE and TE need to verify the decoder runs properly  | Not needed for TE and UE | Needed for both TE and UE |
| Feasibility of test decoder verification procedure | Offline alignment is required to ensure the performance. | Offline alignment is required to ensure the performance. | Not needed for TE and UE | Offline alignment is required to ensure the performance even with partial knowledge of test decoder. |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | May not reflect the actual decoder implemented by gNB vendor | Could reflect the performance in real deployment | May not reflect the actual decoder implemented by gNB vendor (it may also depend on how much similarity exists between the implemented and specified decoder) | May partially reflect the performance in real deployment based on specified parts of test decoder |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Medium. TE needs to cope with multiple decoders from multiple UE vendors | Medium. TE need to implement multiple decoders from different BS vendors | Low | If TE is responsible for the training of test decoder, the requirements for TE is high. If UE or gNB provide the test decoder, the requirements for TE is the same as option 1 and 2. |
| Specification Effort (e.g. test decoder) |  Low |  Low |  High |  High |
| Confidentiality/IP issues | Model exposure is required from UE to TE | Model exposure is required from gNB to TE |  No | No |
| Applicability to different scenarios/conditions/ configurations | Yes | Yes | Conditional Yes. | Conditional Yes. |
| Complexity of actual testing procedure for the ecosystem | Depend on the verification procedure between UE and TE. | Depend on the verification procedure between UE and TE/gNB. | Low | Depend on the verification procedure between UE and TE/gNB. |

**Proposal 4: Take option 1 and 3 as the starting pointing for generating test dataset. How to guarantee the independence of test dataset could be further studied.** |
| [**R4-2315105**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315105.zip) | CATT | **Observation 1: TE vendors probably have to develop different models provided by multiple UE/network vendors and too many models will make the tests unaffordable, if option 1/2 is agreed.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder |   |   | RAN4 **fully** specify the test decoder in spec. | RAN4 **partially** specify the test decoder in spec. |
| Source of decoder training data |  |  | Data set for training can be collected by UE/NW/ the 3rd party.  | Data set for training can be collected by UE/NW/ the 3rd party.  |
| DUT vendor knowledge of the test decoder |   |   | Full knowledge of the test decoder. | Partial knowledge of the test decoder with some unspecified parts.  |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) |   |   | Type 1/2/3. | Type 1/2/3. |
| Test decoder verification procedure at TE and/or DUT |   |   | NA, since the test decoder is fully specified in spec. | May required, since the test decoders provided by different TE vendors could be different. |
| Feasibility of test decoder verification procedure |  |  | NA, since the test decoder is fully specified in spec. | Feasible. |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) |   |   | Depends on the test decoder decided during discussion. | Depends on the test decoder decided during discussion. |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) |   |   | Develop the models fully specified by RAN4 in spec.  | Develop the models partially specified by RAN4 in spec with some development flexibility. |
| Specification Effort (e.g. test decoder) |   |   | Heavy workload to fully specify test decoder. | Less heavy workload to partially specify test decoder. |
| Confidentiality/IP issues |   |   | NA. | NA. |
| Applicability to different scenarios/conditions/ configurations |   |   | Depends on the test model decided during discussion. | Depends on the test model decided during discussion. |
| Complexity of actual testing procedure for the ecosystem |  |  | Relatively easy for the ecosystem. | Relatively easy for the ecosystem. |

**Observation 2: Performance degradation due to model mismatch could happen in the field, if option 1 is agreed.** **Proposal 1: RAN4 can further discuss what aspects of AI/ML models need to be specified in spec in a high level to facilitate comparison, e.g., model structure, activation function, etc.**  |
| [**R4-2315310**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315310.zip) | Qualcomm, Inc. | **Proposal 1: If RAN4 agreed any common assumptions for test decoder/network side models, they are applicable to all the options.****Proposal 2: Decoder/network side model verification procedure is not needed for option 1 and 3 since the DUT vendor or RAN4 specification determines the decoder/network side model. However, for option 2, the decoder/network side model is from a third party other than the DUT vendor and specification group, decoder/network side model verification procedure is needed to ensure the effectiveness of the test.****Proposal 3: Consider at least the following three options for test decoder/network side model in the noted table in the WF:*** **Option 1: test decoder/network side model is provided by the vendor of the DUT**
* **Option 2: test decoder/network side model is provided by the vendor of the decoder/network side model, or any third party other than the DUT vendor**
* **Option 3: The reference decoder(s) are fully specified and captured in RAN4 specifications**

**We are open to discuss option 4 and companies preferring option 4 can clarify what are the differences between option 1/2/3 and option 4.****Besides filling in the entries, we propose to add the following rows to the table:*** **As part of clarification of options: Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration)**
* **As part of pros/cons analysis: Forward compatibility when new AI models are invented**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1: DUT provides decoder** | **Option 2: Decoder not from DUT and Spec** | **Option 3: Full decoder specification in standard** | **Option 4: partially specified decoder** |
| Clarification of options |
| Source of the test decoder  |  DUT vendor |  TBD but not DUT vendor and specification (refer to as the provider) |  RAN4 specification |  Can be option 1 or 2 |
| Source of decoder training data  | DUT vendor generated by its decoder | The provider  | DUT vendor generated by the RAN4 decoder | Can be option 1 or 2, depends on the Source of the test decoder |
| DUT vendor knowledge of the test decoder | Full knowledge based on its design |  Partial knowledge based on common assumptions and data provided by the provider | Full knowledge based on RAN4 spec |  Can be option 1 or 2, depends on the Source of the test decoder |
| Supported training collaboration type (source of training data should be consistent with the collaboration type)*\* Note: RAN4 specification of training collaboration procedure before the test is not needed and simulation assumption on training/collaboration type can be discussed separately in the WI stage.* |  All the training collaboration types are applicable and feasible | Based on the collaborative decision of DUT and decoder provider, outside of 3GPP scope and RAN4 doesn’t have to clarify it from test specification perspective. |  All the training collaboration types are applicable and feasible if it is a sequential training with a fixed decoder |  Can be option 1 or 2, depends on the Source of the test decoder |
| Test decoder verification procedure at TE and/or DUT |  No need since:*No matter the failure is from decoder or DUT, it’s DUT vendor’s responsibility* |  Option A: the verification is neededOption B: the verification is not needed |  No need since:*Any verification procedure can be done before agreeing the decoder* |  Can be option 1 or 2, depends on the Source of the test decoder |
| Feasibility of test decoder verification procedure | N/A | Can discuss after the above row agreed | N/A |  |
| Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | One | Option A: OneOption B: More than oneOption C: RAN4 doesn’t need to make decision | One |  |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) |  DUT may not have full knowledge of the decoder |  Align better with reality, DUT has partial knowledge of the decoder |  DUT may not have full knowledge of the decoder |   |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) |  More flexibility is required to load different models |  More flexibility is required to load different models |  Least flexibility is required |   |
| Specification Effort (e.g. test decoder) |  Low |  Low |  High |   |
| Confidentiality/ IP issues | FFS | FFS |  Decoder is captured in spec, no IP issues |   |
| Applicability to different scenarios/conditions/ configurations |  No differences between options |   No differences between options |   No differences between options |   |
| Complexity of actual testing procedure for the ecosystem | Medium: DUT vendors have to design and deliver the model to TE | High: Decoder provider has to design and deliver the decoder to TE, offline collaboration between decoder provider and DUT vendor, and decoder verification procedure might be required | Low |  |
| Forward compatibility when new AI models are invented | May have forward compatibility if common assumptions doesn’t exclude new models | May have forward compatibility if common assumptions doesn’t exclude new models | Fixed decoder, no forward compatibility |  |

**Proposal 4: Add the following note to the “Supported training collaboration type”: *Note: RAN4 specification of training collaboration procedure before the test is not needed and simulation assumption on training/collaboration type can be discussed separately in the WI stage.*****Proposal 5: Our opinions of the above listed options in clarification of test decoder options are listed below:*** **Issue 1: Whether decoder verification procedure is needed for option 2 => Option A: applicable**

**The decoder/network side model is from a third party other than the DUT vendor and specification group, decoder/network side model verification procedure is needed to ensure the effectiveness of the test*** **Issue 2: Number of test per simulation setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) => Option A: one**

**One test is sufficient per simulation setup given that the encoder/UE side mode and decoder/network side model pair is tailored to each setup by appropriate training procedures.****Proposal 6: Consider complexity range as part of the common assumptions for test decoder in CSI feedback use case.** |
| [**R4-2315324**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315324.zip) | CMCC | ***Proposal 1: it is prefered to define requirements and/or tests to verify the generalization/*** ***scalability of AI/ML.******Proposal 2: it is proposed that the generalization/scalability performance can be verified over various scenarios and/or configurations.*** ***Proposal 3: for 2-sided model testing options, the consideration are provided as following:***

|  |  |  |  |
| --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 |
| Clarification of options |
| Source of the test decoder | vendor of the encoder | vendor of the decoder |  N/A |
| Source of decoder training data |  |  |  |
| DUT vendor knowledge of the test decoder | Need collabaration between DUT vendor and source of the test decoder | Need collabaration between DUT vendor and source of the test decoder |  Full knowledge |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | Type 1 | Type 3 |  N/A |
| Test decoder verification procedure at TE and/or DUT | Needed at TE, and how to verify is FFS | Needed at TE, and how to verify is FFS |  Not needed |
| Feasibility of test decoder verification procedure | Feasible | Feasible | N/A |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | Low, since the test decoder may be mismatch with the decoders deployed in the field, and UE may easily pass the test since UE has full knowledge of the decoder |  High | Low, since the test decoder may be mismatch with the decoders deployed in the field, and UE UE may easily pass the test since UE could train the model based on the specified decoder |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) |  High, since TE may need to implement multiple docoders from different vendors |  High, since TE may need to implement multiple docoders from different vendors |  Low  |
| Specification Effort (e.g. test decoder) | Low, if the decoder is up to implementation, there is no spec impact | Low, if the decoder is up to implementation, there is no spec impact | High, since the decoder(s) are fully specified and captured in RAN4 spec, may results in long discussion  |
| Confidentiality/IP issues |  Need to be condiered |  Need to be condiered | Since the decoder(s) are fully specified and captured in RAN4 spec, no IP issues |
| Applicability to different scenarios/conditions/ configurations | This is pending on how to design the test to guarantee the generalization | This is pending on how to design the test to guarantee the generalization | This is pending on how to design the test to guarantee the generalization |
| Complexity of actual testing procedure for the ecosystem |  |  |  |

 |
| [**R4-2315482**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315482.zip) | Apple | 1. *For Option 3 - Fully specified model means the model and weights are captured in RAN4 specification.*
2. **Confirm that for option 3 model and weights are fully specified in RAN4 specification.**
3. *For Option 4 – It is not clear what partially specified means. Is only the architecture of the model or is the model is specified, but training is left to TE.*
4. **Further clarify what option 4 – partially specified model means. If only architecture of model is specified, or model is specified with training left to TE.**
5. **Update the wording in options 1-4 to use test decoder/encoder as - For UE under test - Option 1: Test decoder is provided by the vendor of the encoder under test so that the encoder and decoder are jointly designed and trained**

**Option 2: Test decoder is provided by the vendor of the decoder(infra-vendors) so that the encoder and decoder are jointly designed and trained****Option 3: The test decoder(s) are fully specified and captured in RAN4 spec to ensure identical implementation across equipment vendors without additional training procedure needed.****Option 4: The test decoder(s) are partially specified and captured in RAN4 spec.For gNB under test - Option 1: Test encoder is provided by the vendor of the decoder under test so that the encoder and decoder are jointly designed and trained****Option 2: Test encoder is provided by the vendor of the encoder(UE) so that the encoder and decoder are jointly designed and trained****Option 3: The test encoders(s) are fully specified and captured in RAN4 spec to ensure identical implementation across equipment vendors without additional training procedure needed.****Option 4: The test encoders(s) are partially specified and captured in RAN4 spec.**Table 1: Options analysis for test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** |
| **Source of the test decoder** |  UE vendor | NW vendor | RAN4 spec | RAN4 spec and TE vendor |
| **Source of decoder training data** | No additional training is needed at TE since UE vendor gives the model after training | No additional training is needed at TE since NW vendor gives the model after training |  If model is fully specified, no additional training is needed | UE vendor provides training data to TE for training.  |
| **DUT vendor knowledge of the test decoder** |  Complete |  No knowledge |  Complete | partial based on what is specified |
| **Supported training collaboration type (source of training data should be consistent with the collaboration type)** |  Type 1 with UE side trainingType 3 training with UE first |  Type 1 with NW side trainingType 3 with NW first training |  This is possibly the parallel training under Type 3 which was not further discussed in RAN1 | Unsure as it is not clear what is partially specified |
| **Test decoder verification procedure at TE and/or DUT** |  Up to the UE to guarantee the decoder works with the encoder. Tested at DUT |  The reference encoder used by NW should be made available to verify the decoder provided by NW |  Should be guaranteed when specified, no need to test |  Unsure |
| **Feasibility of test decoder verification procedure** | Feasible at DUT | Feasible at TE with reference encoder from NW  | During specification effort | Unsure |
| **Pros/Cons analysis** |
| **Reflection on the real deployment (knowledge of model, training type, etc.)** | It could potentially be the same model used in testing and deployment with Type 1 UE side training  |  It could potentially be the same model used in testing and deployment with Type 1 NW side training |  It could limit implementation of encoder model to work with the test decoder, or potentially lead to mismatch in model used for testing and in deployment |  With no knowledge of the full decoder, there might be a mismatch in encoder-decoder pair, as test decoder is only partially specified. It could limit implementation of encoder model to work with the test decoder, or potentially lead to mismatch in model |
| **TE requirements to deploy the decoder (e.g. training, complexity, interoperability)** |  UE vendor provides trained model to TE |  NW vendor provides trained model to TE |  Assume no training needed, fully specified. Complex to ensure inter-operability with different UE vendor encoder models |  Train the model. Complex to ensure inter-operability with different UE vendor encoder models |
| **Specification Effort (e.g. test decoder)** |  Need not be specified |  Need not be specified |  It would take a lot of time and effort to reach consensus on a test decoder model to be specified by RAN4 |   It would take a lot of time and effort to reach consensus on a test decoder model to be specified by RAN4 |
| **Confidentiality/IP issues** |  Disclose to TE | Disclose to TE | decoder already specified, and disclosed |  decoder already specified, and disclosed |
| **Applicability to different scenarios/conditions/ configurations** |  The model should be applicable for the scenarios/ configurations tested for in RAN4 |  The model should be applicable for the scenarios/ configurations tested for in RAN4 |  The model should be applicable for the scenarios/ configurations tested for in RAN4 |  The model should be applicable for the scenarios/ configurations tested for in RAN4 |
| **Complexity of actual testing procedure for the ecosystem** | complex | complex | complex | complex |

 |
| [**R4-2315595**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315595.zip) | ROHDE & SCHWARZ | **Observation 1a**: For options 1 and 2, the usage of an arbitrary decoder may not be feasible, due to complexity constraints. RAN4 should provide an upper bound to the model complexity (e.g., number of parameters or FLOPs per second) that the TE shall support. This observation does not preclude that other complexity measures than the FLOPs per second or the number of parameters may be more relevant to identify the maximum acceptable complexity for the TE.**Observation 1b**: For options 1 and 2, IP issues may arise, because the decoder needs to be available in the TE. This may be particularly relevant in option 2, where the reference decoder must be disclosed by infra-vendors to test the encoder in the UE. Option 1 might lead to potential IP issues for UE vendors, since the decoder (provided in this case by the UE manufacturer) may indirectly yield insights about the encoder.**Observation 1c:** For option 1, the conclusiveness of the test may be limited, because the decoder implementation provided by the UE-vendor may significantly differ from decoders that are deployed in the field by infrastructure vendors.**Observation 2a:** Option 3 involves the usage of a fully specified decoder. We remark that the decoder designs are likely to evolve over time as AI/ML R&D advances. Therefore a “good” decoder at any given time may not be good anymore in a few months. Today e.g. transformer networks are very popular, but other neural network architectures may yield superior results in the months and years to come. The evolution of these models may lead to frequent updates of the RAN4-specified reference decoders, which may result in an increased standardization effort.**Observation 2b:** Option 3 leads to the usage of a specified set of well-defined decoders. However, if only a limited number of decoders is used, the test only proves that the UE can cope with the given decoder(s), but no statement can be given that the UE can act well with different vendors or models in the field. This can lead to a loss of generalisation capability of the employed encoder as the decoder(s) is fixed. Moreover, the usage of a tuned encoder for a very specific decoder may not provide confidence on the actual performance of a UE in the field.**Observation 3a:** Option 4 may provide a trade-off between the reproducibility of the test results and proof of the ability of an encoder to work with decoders by different infrastructure vendors. RAN4 may specify, e.g., a range on the decoder complexity or parts of the decoder model architecture (e.g., the number of layers and type of interconnections). Moreover, part of the specification can involve the employed propagation model (e.g., TDL-A or UMi), such that a TE vendor can train a decoder for the requested scenario. This would shift the focus of the test definition from defining a specific model (which may not prove the UE’s real capabilities) to testing the UE capabilities in a certain environment.**Observation 3b:** Option 4 leaves a certain degree of uncertainty on the actual used decoder in the TE. Therefore, the actual absolute result for the same DUT may vary between different TEs, but the test may still be well defined by formulating an acceptance criterion in terms of relative KPIs. For instance, that the UL throughput by means of an AI model increases by at least X% with respect to a non AI-enhanced case or that the UL throughput is within X% the maximum expected UL throughput. This type of definition is not unusual in conformance specification tests (e.g., 38.521-1).***Proposal 1: A partial specification of the decoder may provide a more realistic evaluation of the UE capabilities. The left-over uncertainty due to the incomplete definition of the decoder may be limited by appropriate formulation of the test acceptance criteria.*** ***RAN4 shall investigate which aspects of the decoder shall be standardized, e.g., model complexity.*** |
| [**R4-2315600**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315600.zip) | Ericsson | [Observation 1 It is not clear whether issue 1-6 and the conclusion is referring to observation of specific models or performance monitoring of the AI functionality.](#_Toc146719172)[Observation 2 Measurement of performance of a UE against RAN4 metrics in the field would be by means of RAN1/RAN2 procedures, not test equipment](#_Toc146719173)[Observation 3 It is not envisaged that compliance testing would be performed in the field. Development of standardized metrics for comparing performance may be of interest, but minimum requirements may not be.](#_Toc146719174)[Observation 4 If measurement reports are provided to the network for monitoring the performance of an AI functionality, RAN4 should consider how accurately the measurement reports would really relate to ground truth.](#_Toc146719175)[Observation 5 Even if the reported information does not correspond with exact accuracy to ground truth, it may still be useful for a coarse metric on model performance.](#_Toc146719176)[Observation 6 If performance monitoring consists of standardized information being periodically sent to the network to assess an assumed metric, the extent of RAN4 involvement should be discussed. It may be for the metric itself, or just the accuracy requirements on the reported information.](#_Toc146719177)[Observation 7 If the UE reports performance or reliability information, RAN4 could set requirements on the minimum accuracy for the performance/reliability report.](#_Toc146719178)[Observation 8 Testing of a requirement on the accuracy of reliability / performance reporting could be done in a test-house as part of compliance testing, since TE could generate ground truth and assess the real reliability/performance.](#_Toc146719179)[Observation 9 It may not be straightforward for a test specification to force different levels of reliability/performance for testing a metric.](#_Toc146719180)[Observation 10 The task in RAN4 is to study how to provide requirements and tests that prove that an AI model can generalize.](#_Toc146719181)[Observation 11 The limits of non-AI algorithm performance when encountering different conditions to the requirement/test conditions are not captured in any specification or TR.](#_Toc146719182)[Observation 12 To perform as well as non-AI, AI models should be as generalizable as non-AI algorithms.](#_Toc146719183)[Observation 13 It is important to understand, for each use case, whether the AI model generalizes smoothly and what level of performance degradation is seen.](#_Toc146719184)[Observation 14 If the generalization behaviour of the model is not smooth, testing under a large variety of short samples of different conditions could be considered.](#_Toc146719185)[Observation 15 Regarding training, there is no need for RAN4 to consider training procedures. However accuracy requirements on training data may be considered if the data is exchanged over standardized interfaces.](#_Toc146719186)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Training | Model monitoring and Model selection/(de)activation/switching/fallback | Model Inference |
| N/W-UE Collaboration Level-x | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | N/A | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-y | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf. | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-z | N/A for one-sided model training(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.)N/A for two-sided model online training and FFS offline training.  | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf.No interoperability aspects for  - model deployment/update/transfer/delivery from/to model storage | Interoperability guaranteed by - Use case KPI |

[Observation 16 For LCM, RAN4 should consider whether requirements on accuracy/reliability of model monitoring reports should be devised.](#_Toc146719187)[Observation 17 Collaboration level x does not seem to provide a means of monitoring or reacting to AI performance. If AI performance characteristics would differ to non-AI, then this may cause inefficiencies in RRM, scheduling and management of UEs.](#_Toc146719188)[Observation 18 For collaboration levels y and z, for LCM there may be a need to consider RRM like requirements on interruption, activation time etc.](#_Toc146719189)[Observation 19 For model inference, there is a need to consider whether the requirement and testing coverage is sufficient to ensure generalization (and/or whether this is possible)](#_Toc146719190)[Observation 20 For model inference, there is a need to consider whether performance should be tested for model updates on representative hardware prior to distribution and how/whether this should be captured in the specification.](#_Toc146719191)Based on the discussion in the previous sections we propose the following:[Proposal 1 Assume that performance monitoring requirements in the field would apply for functionality monitoring.](#_Toc146719192)[Proposal 2 RAN4 discuss further the feasibility of a requirement and test for UE reliability/performance reporting for monitoring of UE sided models.](#_Toc146719193)[Proposal 3 RAN4 get an understanding (e.g., from RAN1) of how graceful the degradation of AI model performance with changing scenario parameters is expected to be for each of the use cases.](#_Toc146719194)[Proposal 4 Where AI degradation is not smooth with changing parameters (for an AI model not trained over a wide enough range of scenarios), RAN4 study possible mechanisms for testing generalizability.](#_Toc146719195)[Proposal 5 For collaboration level x, no LCM requirements on interruption time, activation timer etc. are considered.](#_Toc146719196) |
| [**R4-2315728**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315728.zip) | vivo | ***Observation 1: RAN4 testability study should consider all the relevant parts for defining performance requirements and testing.******Proposal 1: RAN4 to define reference model for defining performance requirements for one-sided model.******Proposal 2: In 2-side model use case, both reference encoder and reference decoder are introduced for defining performance requirements for UE side encoder.******Proposal 3: Fully specified and partially specified options, i.e., option 3 and/or option 4, are used as baseline for RAN4 to specify reference model for defining requirements for different use cases.******Proposal 4: Take into consideration the above analysis on pros/cons/feasibility for different reference decoder/encoder definition.***Table 1 Summary of 4 options for testing of 2-sided model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder |  UE vendor | NW vendor  |  RAN4 specification | Partially specified by RAN4 specification.Further developed by TE/UE/NW vendors |
| Source of decoder training data | * UE vendor
* NW vendors
* RAN4 specification
 | * NW vendors
* RAN4 specification
 | * RAN4 specified
* UE vendors
* NW vendors

Note: only used for developing test decoder | * RAN4 specification
* TE/UE/NW vendors
 |
| DUT vendor knowledge of the test decoder |  Full knowledge | None |  Full knowledge | Full knowledge or partial knowledge |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | * No need to consider training collaboration, or
* Type 1/2/3 training
 | * No need to consider training collaboration, or
* Type 1/2/3 training
 | * No need to consider training collaboration, or
* Type 1/2/3 training
 | * No need to consider training collaboration, or
* Type 1/2/3 training
 |
| Test decoder verification procedure at TE and/or DUT | For DUT: Not neededFor TE: Needed | For DUT: NeededFor TE: Needed | For DUT: Not NeededFor TE: Needed | For DUT: NeededFor TE: Needed |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| Pros/Cons/Feasibility analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | * Impact of the model mismatch could happen since network vendors may use quite different decoder in the field from the test decoder.
 | * Could reflect the performance in the field since network vendors may use same or similar decoder in the field as the test decoder.
* Model transfer may be used to avoid model mismatch in the field, e.g., the corresponding encoder of the test decoder is transferred from NW to UE.
 | * If the test decoder(s) is generated from the well-designed datasets that could reflect real deployment, RAN4 tests could guarantee the AI/ML model performance in the field. However, performance may degrade if low quality datasets are provided.
* The encoder passed the test may not work well for the decoder in the field since the fully specified test decoder may not be implemented in the field. UE may have to implement an additional encoder only for the RAN4 test.
 | * If model is trained with well-designed datasets that could reflect real deployment, RAN4 tests can guarantee the AI/ML model performance in the field. However, performance may degrade if low quality datasets are provided.
* If network/UE vendors consider the partially specified test decoder as reference for implementation, then the performance in real deployment can be ensured.
 |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | TE has to support all the test decoders provided by different UE vendors.The test decoder should be provided in open format. Otherwise, one to many offline co-engineering is needed. | TE has to support all the test decoders provided by different network vendorsThe test decoder should be provided in open format. Otherwise, one to many offline co-engineering is needed. | Easy for TE vendors to implement the fully specified test decoder. | Easy for TE vendors to implement if the partially specified test decoder is further developed by TE vendor.If it is further developed by UE/NW vendors, similar analysis as option 1 or option 2. |
| Specification Effort (e.g. test decoder) | Low: Basic assumptions (scenarios, configurations etc.) | Low: Basic assumptions (scenarios, configurations etc.) | High: Everything (e.g. model structure, model parameters, value of parameters, etc. for different scenarios/configurations/conditions and/or training dataset etc) | Medium to high: key parts (e.g. model structure, model parameters and other necessary part if any, training datasets for different scenarios/configurations/conditions etc.) |
| Confidentiality/IP issues |  FFS |  FFS |  None |  FFS |
| Applicability to different scenarios/conditions/ configurations |  Applicable | Applicable |  Applicable |  Applicable |
| Complexity of actual testing procedure for the ecosystem | * Offline co-engineering between TE vendor and UE vendors may be needed.
* All UE vendors should provide its own test decoder
* How would TE select the corresponding test decoder for a UE under test since it is specific to a UE vendor?
 | * Offline co-engineering between TE vendor and network vendors may be needed.
* Whether should all network vendors provide test decoder?
* How would TE select the corresponding test decoder for a UE under test or would the DUT pass test with all the test decoder from different network vendors?
 | * TE only needs to implement the test decoder
 | * TE needs to train and implement partially specified test decoder, if it is further developed by TE vendors.
* There could be different performance from TE vendors
* Similar analysis as option 1 or option 2, if it is further developed by UE or NW vendors.
 |
| Relationship with reference decoder/encoder for defining requirement | Alt 1: same as reference decoderMay not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.Alt 2: different from reference decoderUE may not pass the tests due to different test decoders are used for defining requirements and tests. | Alt 1: same as reference decoderMay not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.Alt 2: different from reference decoderUE may not pass the tests due to different test decoders are used for defining requirements and tests. | Alt 1: same as reference decoderPossible to define requirements and be able to calibrate results from companies.Alt 2: different from reference decoderThere is no reason to specify test decoder different from that is used for defining requirements. | Alt 1: same as reference decoderThere is good chance that the results among companies can be calibrated as the performance of the model could largely be decided by the specified part.Possible to define requirementsAlt 2: different from reference decoderThere is no reason to specify different test decoder than that is used for defining requirements. |

***Proposal 5:*** ***Take into consideration the summary of 4 options for testing of 2-sided model in Table 1.******Proposal 6: The test decoder/encoder design should aim for testing the encoder/decoder to be used in practical network.******Proposal 7: Reference block diagrams in Fig 3 and Fig 4 for one-sided model and 2-sided model, and functional block description in Table 1 are used for test framework for AI/ML.*****Table 2: Description of reference functional blocks**

|  |  |
| --- | --- |
| Functional block | Description |
| DUT | Device under test. It can be UE or gNB. |
| Test system | A system to test AI/ML functionality/performance. It may be test equipment or gNB in practical NW.  |
| Test setup | Setup test environment based on design of test cases  |
| Data generator | This function is to generate test dataset for the ongoing test. |
| AI/ML model control | In tests for verifying model inference performance, AI/ML model control may be used for model selection, and model activation if necessary.In tests for LCM procedure, AI/ML control may be used for model selection, switch, activation, deactivation, transfer, delivery, update or model monitoring |
| Test model | Test decoder/encoder for UE and gNB, respectively for 2-sided model. |
| Performance requirements verification | This function is to verify if the performance requirements for a test can be met in the ongoing test. |
| LCM requirements verification | This function is to verify if the LCM related requirements for a test can be met in the ongoing test. |

Fig 3. Reference block diagram for one-sided AI/ML modelFig 4. Reference block diagram for 2-sided AI/ML model |
| [**R4-2315754**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315754.zip) | ZTE Corporation | **Observation 1: Level x is implementation-based AI/ML operation without any dedicated AI/ML-specific enhancement (e.g., LCM related signaling, RS) collaboration between network and UE and no collaboration for level x based on RAN1 outcomes.****Proposal 1: RAN4 shall not study the interoperability aspect for level x based on previous meetings in RAN1.****Observation 2: For the level y collaboration, it is clarified as the signaling-based collaboration without model transfer.****Proposal 2: RAN4 needs to consider the interoperability for collaboration level y based on more RAN2 progress since it is the signaling-based collaboration.****Observation 3: Two categories of models including Proprietary-format models and Open-format models were proposed by RAN1. For the Proprietary-format models, due to the lack of inter-operation and recognition between vendors, it is hard to standardize based on the unified specification identification.****Observation 4: Regarding to the Open-format models, the interoperability is feasible.****Proposal 3: RAN4 can focus on the Open-format models firstly, and discuss which core part and performance part requirements should be identified and how to define. On the other side, the test framework and procedure should also be discussed. At the meanwhile, RAN4 needs to wait for RAN1 progress on Open-format models.****Proposal 4: RAN4 shall study the basic structure of the open-format and consider how the common understanding defined between different vendors.****Proposal 5: From the perspective of test, both functionality test and performance test should be considered.****Observation 5: model inference is the core component of AI/ML. Two aspects should be considered to verify: 1) The outputs are the results from the AI/ML inference model rather than the traditional solution; 2) The accuracy of outputs meet the requirement.****Observation 6: the latency requirements and the KPI and the threshold for judging the model performance shall be considered in RAN4.****Observation 7: The test environment is quite different from the field, so the assumptions is a little limitations and the generalization will be degraded.****Observation 8: The options may not work well, and the requirements based on these assumptions may not have significance.** **Observation 9: The TU is limited in RAN4 for AI/ML.****Proposal 6: RAN4 shall study one-sided model firstly and the discussion of two-sided model shall be deprioritized since the two-sided model is only for CSI case.****Observation 10: 3GPP channel models have stable performance and sufficient physical meanings. It is convenient to generate large number of samples using 3GPP channel models.****Proposal 7: In order to guarantee the stable performance and convenience, RAN4 shall study and use the dataset based on TR 38.901 firstly.** |
| [**R4-2315996**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2315996.zip) | Huawei,HiSilicon | ***Proposal 1:*** Table for description of 2-sided model testing options.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder | UE vendor | gNB vendor | RAN4 Spec | RAN4 Spec |
| Source of test decoder training data | Out of RAN4 scope | Out of RAN4 scope | RAN4 Spec | RAN4 Spec |
| DUT vendor knowledge of the test decoder | YES | Out of RAN4 scope | YES | Partially (Whether the unspecified part is known is out of RAN4 scope)  |
| Supported training collaboration type ~~(source of training data should be consistent with the collaboration type)~~ | At least for * Type 1 UE-side joint training
* Type 3 UE-first sequential training
 | At least for* Type 1 NW-side joint training
* Type 3 NW-first sequential training
 | Only if the test decoder can match well with models where the model structure is supported at UE | Depend on which part of the test decoder is specified and which entity provides the unspecified part |
| Test decoder verification procedure at TE and/or DUT | Need | Need | Need | Need |
| Feasibility of test decoder verification procedure | No consensus | No consensus | No consensus | No consensus |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | It depends on training dataset | It depends on training dataset | It depends on training dataset | It depends on training dataset |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | It depends on the total number of the test model  | It depends on the total number of the test model  | It depends on the total number of the test model  | It depends on the total number of the test model  |
| Specification Effort (e.g. test decoder) | * Procedure for verifying the decoder
 | * Procedure for verifying the decoder
 | * Align on assumptions for both for model structure and for model parameters
* It also depends on whether model structure per use case or per configuration/scenario, and whether model parameters per configuration/scenario
 | Depend on which part of the test decoder is specified and which entity provides the unspecified part |
| Confidentiality/IP issues | YES (if model exchange between TE vendor and UE vendor) | YES (if model exchange between TE vendor and NW vendor) | NO  | NO  |
| ~~Applicability to different scenarios/conditions/ configurations~~ |  |  |  |  |
| ~~Complexity of actual testing procedure for the ecosystem~~ |  |  |  |  |

***Proposal 2:*** The interoperability is verified via core requirements and performance requirements. |
| [**R4-2316230**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316230.zip) | OPPO | **Proposal 1: Regarding the testability of two-sided model, should introduce test encoder(s) to collaborate with the NW decoder under test.****Proposal 2: Regarding the testability of two-sided model, should introduce test decoder(s) to collaborate with the UE encoder under test.****Observation 1: Pros and cons for different options on test decoder are shown in table 1.****Table 1 pros and cons of different options for test decoder**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder | UE | NW | RAN4 spec | RAN4 spec and [TE] |
| Source of decoder training data | UE | NW | RAN4 spec | RAN4 spec and [TE] |
| DUT vendor knowledge of the test decoder |  Full knowledge | No, need assisted indication of the test decoder knowledge |  Full knowledge | Partially, may need assisted indication of the test decoder knowledge |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | Type1 and Type3,For example, UE first trains its encoder and a corresponding decoder, and then transmits the decoder to NW, or transmits a dataset that can assist NW in training its decoder to NW. Under these conditions, when testing UE side encoder, the NW side decoder provided by the UE can be used as the test decoder.Type2 is deprioritized in RAN1#111 meeting. | Type1 and Type3,For example, NW first trains its decoder and a corresponding encoder, and then transmits the encoder to UE, or transmits a dataset that can assist UE in training its encoder to UE. Under these conditions, when testing UE side encoder, the NW side decoder provided by the NW can be used as the test decoder.Type2 is deprioritized in RAN1#111 meeting. | Type1 and Type3,Type2 is deprioritized in RAN1#111 meeting. | Type1 and Type3,Type2 is deprioritized in RAN1#111 meeting. |
| Test decoder verification procedure at TE and/or DUT | No consensus | No consensus | No consensus | No consensus |
| Feasibility of test decoder verification procedure | No consensus | No consensus | No consensus | No consensus |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) |  Depends on training data |  Depends on training data |  Depends on training data |  Depends on training data |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | High, TE needs to support various test decoders specific to different devices, particularly when there are a multitude of test decoder providers(e.g. UEs).And also depends on the granularity of test decoder. | Relatively high, TE needs to support different test decoders from NW vendors.And also depends on the granularity of test decoder. | Medium,TE needs to provide support for a limited number of test decoders that specified and captured in RAN4 spec.And also depends on the granularity of test decoder. | Medium, TE needs to provide support for a limited number of test decoders that specified and captured in RAN4 spec.And also depends on the granularity of test decoder. |
| Specification Effort (e.g. test decoder) | Medium,The test may involve test model/data indication and delivery. | Medium,The test may involve test model/data indication and delivery. | High,Consensus of a test model(s) in RAN4 is a challenging task. | High,Consensus of a test model(s) in RAN4 is a challenging task. |
| Confidentiality/IP issues | With test decoder exposure in training collaboration type1.No model exposure in training collaboration type3. | With test decoder exposure in training collaboration type1.No model exposure in training collaboration type3. | No model exposure. |  No model exposure. |
| Applicability to different scenarios/conditions/ configurations | Depends on the granularity of test decoder and the test data set(s) | Depends on the granularity of test decoder and the test data set(s) | Depends on the granularity of test decoder and the test data set(s) | Depends on the granularity of test decoder and the test data set(s) |
| Complexity of actual testing procedure for the ecosystem | No consensus | No consensus | No consensus | No consensus |
| Friendly to STOA(state of the art) model test | Yes | Yes | No | No |
| Whether model transfer/delivery is needed during the test procedure | Yes | Yes | No | No |

**Proposal 3: Further study the pros and cons for test decoder and test encoder in RAN4.****Proposal 4: Dataset based on TR 38.901, e.g. UMa channel, UMi channel, CDL channel, “legacy approach”, should be considered in RAN4 as the starting point.****Proposal 5: Regarding the AI/ML capabilities, following aspects should be considered** **- Definition of basic AI/ML capability and corresponding testing metrics** **- Definition of different AI/ML capability levels and different testing metrics for different levels** **- Dynamic AI/ML capabilities** |
| [**R4-2316394**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316394.zip) | Nokia, Nokia Shanghai Bell | **Observation 1: Option 2 in the current definition restricts the training to happen only at the network side with joint training. However, it will be better if the network side has the UE side data which can be used to train the test decoder better.****Proposal 1: RAN4 to consider creating a new sub-option to Option 2 (e.g., Option 2a) where the decoder provider is still the NW vendor, however the UE vendor shares the training data with the NW vendor for UE-first Type 3 decoder training.****Observation 2: The format of the test decoder model can be a deciding factor on the option to be used. For e.g., Options 1 and 2 might work with a closed format model (E.g., precompiled model) but Options 3 and 4 might require an open format model that can be retrained further if required.****Proposal 2: RAN4 should discuss on the format of the reference model as a part of discussions on the test decoder.****Observation 3: One of the important factors that will impact the various factors of the workflow of the test decoder is the training type that will be agreed at RAN1. For e.g., Type 1 is better suited for Options 1 and 2, and Type 3 with sequential training can be a go to option for Options 3 and 4. This can impact for e.g., the deployment, training, source of the data set of the test decoder.****Proposal 3: RAN4 should follow the training and collaboration type that is agreed in RAN1 and base its decision of the test decoder workflow on what is concluded in RAN1.****Observation 4: In the case of options 1 and 2: Since both the encoder and the decoder are provided by the DUT vendor, it can be considered as a closed loop DUT vendor-controlled test. This kind of testing will not leave scope for any kind of interoperability of the encoder with different decoders. It will be a difficult task at RAN4 to ensure that the encoder will have reasonable performance with different decoders.****Proposal 4: In case of Options 1 and 2, it needs to be clarified how the test decoder providers can bring in the much-needed interoperability between different encoders provided****Observation 5: In case of options 1 and 2: even though it is a closed loop testing, some agreement still needs to be done between the TE vendors and the DUT-UE vendors. E.g., The Hardware capability required at the TE, complexity of the model, interoperability of the decoder across TE vendors to name a few. Because of this, it will be difficult to completely do away with the specification effort.****Proposal 5: In the case of options 1 and 2, it needs to be discussed at RAN4 on the specification effort that may be required to specify dependencies like hardware capability of the TE, complexity of the model, interoperability of the test decoder across TE vendors etc.,****Observation 6: In case of options 1, 2 and 3: Since the design of the decoder is known to the provider of the encoder, it will be easy to tune the encoder to validate the performance requirement of the AI/ML enabled functionality. Consequently, it would be difficult to guarantee a reasonable performance with the different (other than the test decoder) decoders in the field.****Observation 7: In case of option 3, complete specification of the RAN4 can be a daunting task involving lengthy discussion to align all the stake holders. Also specifying a model can be a complex task with lot of parameters and options to be specified.****Observation 8: In case of option 3, the models can easily get outdated in quick time due to rapid advancement in the AI/ML related technologies and would add significant amount of work in RAN4 for maintenance and too frequent updates. This can be handled with option 4 where the updates can still be handled with the unspecified part of the test decoder.****Proposal 6: We prefer using option 4 as a baseline for the definition of test encoder/decoder.****Observation 9: In case of option 4, the term “partially specified” is very generic. There can be lot of possible options that can be specified. E.g., Model weights, architecture, structure etc.,****Proposal 7: In case of option 4, the definition of partial specification needs to be clarified including what should be specified and what can be left to vendor implementation. For e.g., architecture, structure of the model, model weights, levels of fine-tuning.****Proposal 8: RAN4 to adopt reference testing diagrams for AI/ML functionalities which are based on the current ‘traditional’ test diagrams and include only the strictly needed modifications to capture the nature of the AI/ML-enabled features being tested.****Observation 11: The UE 1-sided and UE-part of 2-sided use cases do not require significantly different minimum performance test setups assuming the corresponding model alignment and potential training/validation are not part of the performance testing procedures, i.e., they are part of initial conditions or test preparation phase.****Observation 12: In the reference test diagrams for AI/ML functionalities it is necessary to indicate separately DL and UL signal paths, such that the setup is not physically different from the traditional test setup(s).****Proposal 9: RAN4 to adopt as reference test diagrams for minimum performance testing of UE 1-sided use cases the setup depicted in Figure 1.**A diagram with text and black arrows  Description automatically generated with medium confidenceFigure 1: Simplified reference testing diagram for UE-side ML Functionalities in 1-sided use cases.**Proposal 10: RAN4 to adopt as reference test diagrams for minimum performance testing of UE-part 2-sided use cases the setup depicted in Figure 2.**A diagram of a software company  Description automatically generated with medium confidenceFigure 2: Simplified reference testing diagram for UE-part ML Functionalities in 2-sided use cases.**Proposal 11: RAN4 to treat the test requirements for functionality-based LCM procedures as core requirements.****Proposal 12: RAN4 to adopt as reference test diagram for Functionality-based LCM procedures for UE 1-sided use cases the setup depicted in Figure 3.**A diagram of a software system  Description automatically generated with medium confidenceFigure 3 – Simplified reference testing diagram for Functionality based LCM procedures for UE 1-sided use cases.**Observation 13: Based on the discussions and agreements in RAN1 and RAN2, it is more natural for RAN4 to first study the ML-enabled Functionality/Feature related inter-operability aspects, and only later address the ML model related aspects if needed.****Proposal 13: RAN4 to adopt and capture in the TR the interoperability analysis Table 2 with focus on ML-enabled Functionality/Feature related aspects for Release 18 use cases.****Observation 14: A special channel model is needed instead of just TDL models that are currently used. More advanced models can be considered for the test such as CDL. Another aspect is time evolution of the channel model.****Proposal 14: RAN4 to study whether TDL models are sufficient for the performance evaluation of AI/ML Enabled CSI feedback use-cases.** |
| [**R4-2316469**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316469.zip) | MediaTek inc. | Table 1. Summary of test decoder design options for 2-sided models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder | vendor of the encoder | vendor of the decoder (infra-vendors) | RAN4 specification (could be TE vender) | Could be vender of encoder, vender of decoder or TE vender  |
| Source of decoder training data | vendor of the encoder | vendor of the decoder (infra-vendors) | without additional training procedure needed | Could be vender of encoder, vender of decoder or TE vender |
| DUT vendor knowledge of the test decoder | Full knowledge of the test decoder | Only from agreed common assumptions | Full knowledge of the test decoder | Depends on source of the test decoder |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | collaboration Type 1 and 3 | collaboration Type 2 and Type 3 |  No need | Depends on source of the test decoder |
| Test decoder verification procedure at TE and/or DUT | No need |  Needed |  No need | Depends on source of the test decoder |
| Feasibility of test decoder verification procedure | NA | Need offline alignment to guarantee the performance | NA | Depends on source of the test decoder |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | May not reflect the real deployment when decoder is implemented in gNB side  | Could reflect the performance in real deployment | RAN4 should study how to guarantee the test decoder can reflect the decoder used in real field | Depends on source of the test decoder |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | TE must support various test decoders from vendors of the encoders | TE must support various test decoders from vendors of the decoders | TE follows the decoder defined in RAN4 specification | Depends on source of the test decoder |
| Specification Effort (e.g. test decoder) | little effortMay need some limit on test decoder to ensure it can be implemented on TE side | little effortMay need some limit on test decoder to ensure it can be implemented on TE side | Consensus of a reference model in RAN4 may be a challenging task | Depends on source of the test decoder |
| Confidentiality/IP issues | Decoder structure will be shared to TE vendor | Decoder structure will be shared to TE vendor |  No issue | Depends on source of the test decoderNo issue if the test decoder is from TE vendor |
| Applicability to different scenarios/conditions/ configurations |  Applicable |  Applicable |  NA |  NA |
| Complexity of actual testing procedure for the ecosystem | HighTE must support various test decoders from vendor of the encoderThe test may involve model delivery | HighTE must support various test decoders from vendor of the decoderThe test may involve model delivery | LowNo model delivery involved if TE implements the test decoder according to the specification | Depends on source of the test decoder |

***Observation 1***: It is not clear that how the test decoder is “partially specified and captured in RAN4 spec”.***Observation 2***: Even if the test decoder is “partially specified” in the RAN4 spec. It is still unclear which entity should provide test decoder, vender of encoder, vendor of decoder or TE vendor. ***Proposal 1***: Remove Option 4 and only discuss option 1, 2 and 3 for 2-sided model. |
| [**R4-2316598**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316598.zip) | Samsung | *Reference block diagrams for testing***Proposal 1: RAN4 shall firstly discuss and agree on the following principles to draft the reference block diagram for 1-sided model and 2-sided model:****- shall NOT contain the block for training;****- shall contain the blocks for model/functionality monitoring and selection/switching/(de)activation/ fallback in DUT;****- shall contain the AI/ML LCM procedure verification and model control in TE;****- shall contain the test scenario generator to enable testing in different scenarios, used for generalization verification aspects.** **Proposal 2: The purpose of introducing the diagram (to be captured in TR38.843) is to derive the potential testing procedure and used as the basis to judge whether certain performance metric is testable, for each use case for normative work.****Proposal 3: RAN4 shall include the following reference block diagram in TR for testing 1-sided model (UE as DUT).** DUT / UE**Fig. 1: Reference block diagram for testing 1-sided model (UE as DUT)****Proposal 4: RAN4 shall include the following reference block diagram in TR for testing the UE-side model of the 2-sided model (based on the example use case of CSI compression).** **Fig. 2: Reference block diagram for testing UE-side model of the 2-sided model (based on the example use case of CSI compression)****Proposal 5: Before defining reference block diagram for testing gNB-side model of the 2-sided model, the test metric and procedure shall be clarified for feasibility.** **Proposal 6: FFS the feasibility of using NR air interface to test either 1-sided model implemented in gNB side or gNB-side model of 2-sided model. If not confirmed, gNB-side model shall be precluded for testing in RAN4.** *Two-sided model framework***Proposal 7: The following clarification of options are provided for option 1-4 test decoder for 2-sided model.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options |
| Source of the test decoder | **The vendor of the encoder/DUT**(i.e., UE vendor) | **The vendor of the decoder**(i.e., BS vendor) | **3GPP specification**(Fully specified) | **3GPP specification**(Partially specified) and model retuning by TE vendors |
| Source of decoder training data | **Depends on UE vendor** | **Depends on BS vendor** | **Depends on 3GPP standardization discussion** | **Depends on 3GPP standardization discussion and vendor input** |
| DUT vendor knowledge of the test decoder | **Full knowledge** | **No** | **Full knowledge** | **Partial knowledge** |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | **Type 1** (Joint training of encoder/decoder at UE-sided) | **Not applicable**(if test decoder is not provided to UE vendors for encoder design)**Or** **Partially Type 2 or 3**(Only if gradient results or test decoder can be provided to UE vendors for encoder design; “Partially” because test decoder is fixed and no further refinement) | **Partially Type 2 or 3**(“Partially” because test decoder is fixed and no further refinement) | **Maybe Type 3 starting at NW side**(if Type 3 collaboration procedure is specified and followed by TE/DUT vendors) |
| Test decoder verification procedure at TE and/or DUT | **No**(Test decoder adopted by TE directly)  | **No and questionable feasibility** (Test decoder adopted by TE directly) | **No**(3GPP-specified test decoder leads to same TE implementation)  | **No** (3GPP-partially-specified test decoder and retuned by TE vendor) |
| Feasibility of test decoder verification procedure | **Not applicable** | **Not applicable**(Question about the how test decoder verification procedure can be made)  | **Not applicable** | **Not applicable** |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | **No**(Can’t reflect real deployment since no evidence shown that BS vendors will adopt decoder provided by UE vendors) | **Yes or Maybe**(Depends on test decoder can be provided to UE for model design) | **Maybe**(Depends on whether specified test decoder can reflect decoder in the field) | **Maybe**(Depends on whether specified test decoder can reflect decoder in the field) |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | **Limited effort** (Model complexity assumption should be aligned by TE and UE vendors) | **Limited effort** (Model complexity assumption should be aligned by TE and BS vendors) | **No effort**  | **No effort**  |
| Specification Effort (e.g. test decoder) | **No specification effort required for test decoder** | **No specification effort required for test decoder** | **RAN4 effort required on specifying test decoder** (including model structure and dataset for training etc.) | **RAN4 effort required on specifying test decoder** (including model structure and dataset for training etc.) |
| Confidentiality/IP issues | **Yes**(Disclosure of UE vendor designed IP) | **Yes**(Disclosure of BS vendor designed IP) | **No issues identified** | **Maybe**(Disclosure of TE vendor designed IP, but depends on model retuning procedure) |
| Applicability to different scenarios/conditions/ configurations | **Yes**, if UE vendors can provide different test decoders accordingly | **Yes**, if BS vendors can provide different test decoders accordingly | **Yes**, if 3GPP can specify different test decoders accordingly | **Yes**, if 3GPP can specify different test decoders and/or TE vendor retune the model accordingly |
| Complexity of actual testing procedure for the ecosystem | **Low**(DUT is only required to be tested against the specified test decoder) | **High**(Conformance tests could be not available since different test decoders from different BS vendors) | **Low**(DUT is only required to be tested against the specified test decoder) | **Low/Medium**(DUT is only required to be tested against the partially specified test decoder, but refinement procedure maybe required for TE) |

**Observation 1: For the reference decoder for test implementation for two-sided models in the UE performance tests, the feasibility of the offline training to obtain UE encoder can be confirmed at least for Option 1 and 3.** **Observation 2: Only Type-1 and Type-3 training collaboration with the offline training manner needs to be considered in Rel-18 RAN4 study on the methodology to obtain the reference model for two-sided model test implementation.** **Observation 3: For the reference decoder to be used in the test implementation for two-sided models for the UE performance tests:**  **- Option 1 can be regarded to match with Type-1 training collaboration, i.e., decoder developed by UE vendors shall be provided to and used by BS vendors directly;** **- Option 2 (if test decoder can be provided to UE for model design) can be regarded to match with partially Type-3 training collaboration, i.e., decoder is provided by gNB vendors for UE-side training, but without further gNB-side training based on labeled data.**  **- Option 3 can be regarded to match with partially Type-3 training collaboration, i.e., decoder is provided UE-side training, but without further gNB-side training based on labeled data.**  **- Option 4 can be regarded to match with Type-3 training collaboration, if the procedure of Type 3 collaboration can be followed by TE and UE vendors.** **Proposal 8: For the test decoder used in the test implementation for two-sided models for the purpose of UE conformance tests, the following modified Option 3 is preferred:** **- (Modified) Option 3: The test decoders are fully specified and captured in RAN4 spec, which are specified corresponding to certain CSI conditions, to ensure identical implementation across equipment vendors without additional training procedure needed.****Proposal 9: If the modified Option 3 is not adopted, RAN4 shall drop all discussion on 2-sided model related topic in Rel-18 SI.** *Test data generation***Proposal 10: FFS on pros and cons for Option-a (dataset provided by 3GPP) and Option-c (methodology provided by 3GPP):**  **- Option-a: Whether the dataset is representative enough is no longer be a problem, but 3GPP have not yet provided a dataset for testing before.**  **- Option-c: For a complex test environment used for AI/ML performance testing, it is possible the test environment (especially for a test within the reasonable test duration) cannot be representative enough due to test limitation (e.g., limited test duration), which can be a problem for repeatability of conformance testing.** *Interoperability aspects***Proposal 11: The interoperability analysis for AI/ML operation for NR air interface are summarized as below, which shall be captured in TR.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Training | Model monitoring and Model selection/(de)activation/switching/fallback | Model Inference |
| N/W-UE Collaboration Level-x | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | N/A | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-y | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf. | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-z | N/A for one-sided model training(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.)N/A for two-sided model online training and FFS offline training.  | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf.No interoperability aspects for  - model deployment/update/transfer/delivery from/to model storage | Interoperability guaranteed by - Use case KPI |

 |
| [**R4-2316855**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_108bis/Docs/R4-2316855.zip) | Keysight Technologies UK Ltd | Table for description of 2-sided model testing options

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** |
| **Source of the test decoder** |  DUT vendor |  NW vendor |  Fully specified by RAN4 |  RAN4 partial specification + further development by TE vendors |
| **Source of decoder training data** |  DUT vendor |  NW vendor | Multiple companies during RAN4 specification process |  RAN4 specification or TE vendors |
| **DUT vendor knowledge of the test decoder** |  Known |  Unknown |  Known |  Partially known |
| **Supported training collaboration type (source of training data should be consistent with the collaboration type)** |  Type 1 (Joint training of the two-sided model at a single side/entity- UE) | Type 3: Separate training at network side and UE sideType 2: Joint training of the two-sided model at network side and UE side respectively 🡪De-prioritized in Rel-18? |  During RAN4 specification process any of the training collaboration type might be used:Type 1: Joint training of the two-sided model at a single side/entityType 3: Separate training at network side and UE side Type 2: Joint training of the two-sided model at network side and UE side respectively 🡪De-prioritized in Rel-18? |  After RAN4 partial specification, TE vendor will need to perform the appropriate training:Type 3: Separate training at network side and UE side  |
| **Test decoder verification procedure at TE and/or DUT** | Not needed for DUTNeeded for TE | Recommended for DUTNeeded for TE | Recommended for DUTNeeded for TE | Recommended for DUTNot Needed for TE |
| **Feasibility of test decoder verification procedure** |  For TE, not feasible unless DUT vendor share encoder, test decoder and data format to shape test data  |  For DUT & TE, not feasible unless NW vendor share test decoder and data format to shape test data |  Feasible, information will be available from RAN4 assumptions and conclusions. |  For DUT vendor, not feasible unless TE vendor share test decoder and data format to shape test data |
| **Pros/Cons analysis** |
| **Reflection on the real deployment (knowledge of model, training type, etc.)** |  Unless collaboration level Z used, there is a risk to have some model mismatch between real deployment and the models tested. Additionally, if scenarios are different from the scenarios considered during the training (generalization), there is a risk of performance gap. |  This option minimizes the mismatch between real deployment and models tested. However, if scenarios are different from the scenarios considered during the training (generalization), there is a risk of performance gap. | Assuming dataset used to produce RAN4 specified test decoder considers information generic enough to represent real deployments, test model defined by RAN4 should ensure acceptable AI/ML performance in the field.Model mismatch between test decoder and AI/ML decoder used by actual NW network will ultimately depend on whether NW vendors used test decoder as baseline for their implementations. If scenarios are different from the scenarios considered during the training (generalization), there is a risk of performance gap. | Assuming dataset used to produce RAN4 partially specified + TE vendors further developed test decoder considers information generic enough to represent real deployments, test model finally defined by TE vendor should ensure acceptable AI/ML performance in the field.Model mismatch between test decoder and AI/ML decoder used by actual NW network is higher than in Option 3 unless TE vendors share the test decoder and data as baseline for their implementations. If scenarios are different from the scenarios considered during the training (generalization), there is a risk of performance gap. |
| **TE requirements to deploy the decoder (e.g. training, complexity, interoperability)** |  TE will need to support a wide range of architectures/interfaces/algorithms (at least one per UE vendor).TE Computational resources requirements should be defined.No additional training required by TE vendor. | TE will need to support a considerable range of architectures/interfaces/algorithms (at least one per infra vendor). TE Computational resources requirements should be defined.No additional training required by TE vendor. |  Single implementation required from TE vendor (just the fully specified test decoder in RAN4)Training will be required during RAN4 specification of the test decoder. |  Single implementation required from TE vendor (although requirements are required for the portion not defined in RAN4: training and final complexity needs to be taken into account). TE vendor will be responsible for designing and training final test decoder. |
| **Specification Effort (e.g. test decoder)** |  Low |  Low |  High |  Medium-high |
| **Confidentiality/IP issues** | DUT vendors disclosing test decoders might reveal indirectly details on their encoders, losing means to differentiate from competitors.Depending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing |  NW vendors disclosing test decoders might reveal indirectly details on their decoders, losing means to differentiate from competitors.Depending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing |  Depending on the source of data used for training the model to be specified, there might be confidentiality issues in this option.  |   Depending on the source of data used for training the model to be specified, there might be confidentiality issues in this option. |
| **Applicability to different scenarios/conditions/ configurations** |  Applicable up to the extend defined by DUT vendor. |  Applicable up to the extend defined by NW vendor. |  RAN4 specification shall assume that test decoder is applicable to different scenarios/conditions/configurations. |  RAN4 specification + further development from TE vendor shall assume that test decoder is applicable to different scenarios/conditions/configurations. |
| **Complexity of actual testing procedure for the ecosystem** |  Potentially, for each DUT, TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test.When executing test, DUT vendor will need to make a manufacturer declaration indicating the test decoder(s) they want to be tested against and for which scenarios (only one or more than one if the DUT is using different AI/ML models for different scenarios). |  DUT will need to be tested against one or multiple test decoders provided by different NW vendors (manufacturer declaration?) Potentially, for each NW test decoder (or even test decoder update?), TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test. |  All DUT will be tested against a fixed set of test decoder(s): the one(s) specified by RAN4.No additional TE integration required once initial implementation of the test system is completed. |  All DUT are supposed be tested against equivalent TE vendor implementation of the test decoder (only one).No additional TE integration required once initial implementation of the test system is completed. |

 |

## Open issues summary

The open issues were grouped in the following sub-topics for further discussion:

1. Test Encoder/decoder option 4 – partially specified
2. Test encoder/decoder options table
3. Reference block diagram for 1 sided model
4. Reference block diagram for 2-sided model
5. Interoperability aspects
6. Channel models

### Sub-topic 3-1

*Test Encoder/decoder option 4 – partially specified*

Some companies brought up the issue that option 4 needs further clarification in terms of what needs to be specified to see if this option is feasible and be able to compare with the other options. Several companies bring up potential benefits of this option so more details need to be understood.

**Issue 3-1: Test encoder/decoder option 4**

* Proposals
	+ Option 1: RAN4 specifies a set of requirements/conditions based on which TE vendors can implement the decoder
		- FFS what requirements/conditions are needed – examples below to be further discussed
			* Range on the decoder complexity or parts of the decoder model architecture like the number of layers and type of interconnections
			* Data to be used for training – propagation model, etc
			* Other parameters/functions/KPIs
		- Performance variation between different TE vendors expected to be small
	+ Option 2: RAN4 specifies a set of requirements/conditions, TE vendor to develop the decoder based on further alignment/collaboration with vendors
	+ Option 3: RAN4 specifies model structure with model parameters based on which UE, NW, or TE vendors can further develop and finalize the test decoder.
		- Model structure is up to RAN1 understanding. An LS can be considered if further clarification is needed.
	+ Option 4: Others
* Recommended WF

To be discussed. Companies are invited to provide proposals on what other parameters/functions/KPIs could be specified

### Sub-topic 3-2

*Test encoder/decoder options table*

**Issue 3-2: Test encoder/decoder options comparison table**

* Proposals
	+ Moderator to provide a table based on inputs from companies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1: DUT provides decoder** | **Option 2: Decoder not from DUT and Spec** | **Option 3: Full decoder specification in standard** | **Option 4: partially specified decoder** |
| Clarification of options |
| Source of the test decoder  |  DUT vendor | Decoder vendor (infra vendor in case of testing UEs)  |  RAN4 specification |  TE vendor, decoder implementation based on RAN4 specifications |
| Source of decoder training data  | Up to DUT implementer | Up to decoder implementer (infra vendor) FFS whether coordination with encoder vendor is required | Not needed, decoder fully specified | 1. RAN4 specifications
2. Up to TE implementation, alignment with UE/gNB vendors might be required
3. UE/gNB vendors
 |
| DUT vendor knowledge of the test decoder | Full knowledge | No knowledge or partial knowledge based on alignment with infra vendors or specifications (RAN4 might specify some high level parameters for the decoder like maximum processing allowed, etc) | Full knowledge based on the specifications | Partial knowledge – depending on what is specified |
| Supported training collaboration type (source of training data should be consistent with the collaboration type)*\* Note: RAN4 specification of training collaboration procedure before the test is not needed and simulation assumption on training/collaboration type can be discussed separately in the WI stage.* | Type 1 and 3Type 2 | Type 3Type 1Type2 depending on offline alignment between DUT vendor and decoder implementer | Type 1, 2, 3 | Type 3Type 1, 2 |
| Test decoder verification procedure at TE and/or DUT | 1. Not needed
2. Needed for TE side
 | Needed (it needs to be proved that the decoder works such that test failure is because of DUT) | Not needed (Test validation would follow the typical RAN5/testing procedures) | Not needed (Test validation would follow the typical RAN5/testing procedures)Needed |
| Feasibility of test decoder verification procedure | ?? | Feasible if network vendor provides a test encoder and passes the testOther options? | Feasible | Feasible if TE vendors shares test decoder  |
| Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | [One] | [Option A: OneOption B: More than oneOption C: RAN4 doesn’t need to make decision] | [One] |  |
| Pros/Cons analysis |
| Reflection on the real deployment (knowledge of model, training type, etc.) | Low1. There could be large performance mismatch with field performance due to mismatch between test decoder and field decoder implemented by gNB
2. Depends on the training data set
 | HighModel which is similar with the test model could be used in the actual deployment by gNBs | LowThere could be large performance mismatch with field performance due to mismatch between test decoder and field decoder implemented by gNB | Medium/Low1. The test decoder may have a large mismatch with the decoders deployed in the field, and UE may easily pass the test since UE could train the model based on the specified decoder
2. May partially reflect the performance in real deployment based on specified parts of test decoder
3. Depends on the data sets used for training
 |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | 1. TE will need to support a wide range of architectures/interfaces/algorithms (at least one per UE vendor). TE Computational resources requirements should be defined. No additional training required by TE vendor.
2. Medium. TE needs to cope with multiple decoders from multiple UE vendors
 | 1. TE will need to support a considerable range of architectures/interfaces/algorithms (at least one per infra vendor). TE Computational resources requirements should be defined. No additional training required by TE vendor.
2. Medium. TE need to implement multiple decoders from different BS vendors
3. High
 | LowLeast complexity required on the TE side | MediumIf TE is responsible for the training of test decoder, the requirements for TE is high. If UE or gNB provide the test decoder, the requirements for TE is the same as option 1 and 2. |
| Specification Effort (e.g. test decoder) | LowSome conditions on the test decoder might be needed to ensure it can be implemented by TE | LowSome conditions on the test decoder might be needed to ensure it can be implemented by TE | HighLong/complicated discussion expected in RAN4 to derive the test decoder | Medium to highLess heavy workload than Option 3Long/complicated discussion expected in RAN4 to agree on the test decoder to be specified in RAN4 |
| Confidentiality/ IP issues | FFSNeed to be consideredDUT vendor might have to expose some aspects of the design to the TE vendorDepending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing | FFSNeed to be consideredDecoder vendor might have to expose some aspects of the design to the TE vendorDepending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing | NoneDecoder is fully captured in the specifications | 1. None
2. Depending on the source of data used for training the model to be specified, there might be confidentiality issues in this option.
 |
| Applicability to different scenarios/conditions/ configurations | Maybe1. pending on how to design the test to guarantee the generalization
2. The model should be applicable for the scenarios/ configurations tested for in RAN4
 | Maybe1. pending on how to design the test to guarantee the generalization
2. The model should be applicable for the scenarios/ configurations tested for in RAN4
 | Maybe1. pending on how to design the test to guarantee the generalization
2. The model should be applicable for the scenarios/ configurations tested for in RAN4

  | Maybe1. pending on how to design the test to guarantee the generalization
2. The model should be applicable for the scenarios/ configurations tested for in RAN4
 |
| Complexity of actual testing procedure for the ecosystem | Medium/HighPotentially, for each DUT, TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test. When executing test, DUT vendor will need to make a manufacturer declaration indicating the test decoder(s) they want to be tested against and for which scenarios (only one or more than one if the DUT is using different AI/ML models for different scenarios). | HighDUT will need to be tested against one or multiple test decoders provided by different NW vendors (manufacturer declaration?) Potentially, for each NW test decoder (or even test decoder update?), TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test. | LowNo additional TE integration required once initial implementation of the test system is completed. | Low/MediumAll DUT are supposed be tested against equivalent TE vendor implementation of the test decoder (only one).No additional TE integration required once initial implementation of the test system is completed. |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes | Yes | No | No/MaybeYes |
| Relationship with reference decoder/encoder for defining requirement | [Alt 1: same as reference decoderMay not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.Alt 2: different from reference decoderUE may not pass the tests due to different test decoders are used for defining requirements and tests.] | [Alt 1: same as reference decoderMay not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.Alt 2: different from reference decoderUE may not pass the tests due to different test decoders are used for defining requirements and tests.] | [Alt 1: same as reference decoderPossible to define requirements and be able to calibrate results from companies.Alt 2: different from reference decoderThere is no reason to specify test decoder different from that is used for defining requirements. ] | [Alt 1: same as reference decoderThere is good chance that the results among companies can be calibrated as the performance of the model could largely be decided by the specified part.Possible to define requirementsAlt 2: different from reference decoderThere is no reason to specify different test decoder than that is used for defining requirements.] |
| Whether model transfer/delivery is needed during the test procedure | [Yes] | [Yes] | [No] | [No] |

* Recommended WF
	+ To be discussed

### Sub-topic 3-3

*Reference block diagram for 1-sided model*

A reference block diagram was discussed in previous meetings, there are several proposals for a diagram that should be captured in the TR

**Issue 3-3: Reference block diagram for 1-sided model**

* Proposals
	+ Option 1: The purpose of introducing the diagram (to be captured in TR38.843) is to derive the potential testing procedure and used as the basis to judge whether certain performance metric is testable, for each use case for normative work. Further discuss the block diagram based on the proposal below (R4-2315066)



* + Option 2: Block diagram is not needed
* Recommended WF
	+ Option 1, details to be further refined

Please provide comments on any changes/clarifications that should be made to the diagram

### Sub-topic 3-4

*Reference block diagram for 2-sided model*

**Issue 3-4: Reference block diagram for 2-sided model**

* Proposals
	+ Option 1: The purpose of introducing the diagram (to be captured in TR38.843) is to derive the potential testing procedure and used as the basis to judge whether certain performance metric is testable, for each use case for normative work. Further discuss the block diagram based on the proposal below (R4-2315066)



* + Option 2: Block diagram is not needed
* Recommended WF
	+ To be discussed

Please provide comments on any changes/clarifications that should be made

### Sub-topic 3-5

*Interoperability aspects*

**Issue 3-5: Interoperability aspects**

* Proposals
	+ Option 1:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Training | Model monitoring and Model selection/(de)activation/switching/fallback | Model Inference |
| N/W-UE Collaboration Level-x | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | N/A | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-y | N/A(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf. | Interoperability guaranteed by - Use case KPI |
| N/W-UE Collaboration Level-z | N/A for one-sided model training(training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.)N/A for two-sided model online training and FFS offline training.  | Interoperability guaranteed by - Model monitoring perf. - Model selection/(de)activation/switching/fallback perf.No interoperability aspects for  - model deployment/update/transfer/delivery from/to model storage | Interoperability guaranteed by - Use case KPI |

* + Option 2: Table is not needed
* Recommended WF
	+ Discuss whether this should be captured in the TR and what changes are needed, if any

### Sub-topic 3-6

*Channel models for testing*

**Issue 3-6: Channel Models for testing**

* Proposals
	+ Option 1: Dataset based on TR 38.901, e.g. UMa channel, UMi channel – “legacy approach”
	+ Option 2: RAN4 should develop more complex models like CDL
	+ Option 3: Use field data – to be further clarified how this data is generated
	+ Option 4: postpone this discussion for now
* Recommended WF
	+ To be discussed