**3GPP TSG-RAN WG1 Meeting #110 R4-2401100**

**Athens, Greece, February 26th – March 1st, 2024**

**Agenda item:** 11.1.5

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [110][141] NR\_AIML\_air

**Document for:** Information

# Introduction

This is the summary thread for issues related to NR AI/ML study in RAN4. The Rel-19 WID was agreed in RP-234039. This is a continuation of the Rel-19 SI, the TP summarizing the RAN4 agreements was approved in R4-23

# Topic #1: General aspects

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2400090**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400090.zip) | CATT | **Proposal 1: For verifying DUT’s AI/ML capability, trained AI/ML models can be defined in RAN4 spec for different use case tests. The following contents can be considered:**  **- This requires that DUT is capable of loading trained AI/ML models.**  **- The RAN4-defined AI/ML models can have classical and simple structures with parameters given.**  **- Discussion on details of RAN4-defined AI/ML models can be combined with that on 2-sided model test to avoid parallel discussion.**  **Proposal 2: Performance requirements can be defined for RAN4-defined AI/ML models. The DUT is confirmed to successfully pass the tests if the performance requirements are met.**  **Proposal 3: For verifying performance gain of AI/ML models/functionalities, RAN4 can define multiple independent** **test cases with different scenarios and configurations which at least include:**  **- Propagation conditions, e.g., channel modes defined for different scenarios (Uma, Umi and InF, etc.) in TR38.901, and UE position (outdoor/indoor), etc.**  **- Configurations considered by RAN1, e.g., bandwidth, carrier frequency, UE speed.**  **The passing conditions for Option 2 can be FFS.**  **Proposal 4: It is necessary to define the test cases with non-static scenarios/configurations to verify the real DUT performance as much as possible, since the wireless environments are changing frequently in the real deployments. Gradually changing configurations and scenario-related parameters can be considered.**  **Proposal 5: If non-static scenarios/configurations are supported, it can also be considered in generalization tests.**  **Proposal 6: Core requirements for LCM related procedures can be define after RAN1/2 makes more progress on procedures and signallings design. The principles of defining core requirements achieved in Rel-18 should be stuck to.** |
| [**R4-2400133**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400133.zip) | CAICT | **Issue 1-5: Latency requirements for data collection or inference**   * Proposals   + Option 1: Latency requirements of data collection for model inference and monitoring could be considered per use case , subject to output from RAN1/2. To be further discussed during WI as needed   + Option 2: RAN4 to define delay requirements for data collection when data are transferred between different entities for inference or monitoring. Similar delay definition in TS 38.133 can be referred, e.g., delay is the period from the moment when data report is triggered to the moment when the entity successfully receives the reported data. Details are FFS when data collection procedure is defined.   + Option 3: Consider data collection latency requirements only for inference and monitoring     - * 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;  t2-1: the processing time for raw input data  t2-2: the measurement time for input data  t3 denotes the uncertainty time;  t4 denotes the propagation delay from UE to report the measurement result to gNB;  t5 denotes the processing time for results ;  N denotes the number of selected samples.   * + Option 4: Do not specify any latency requirements in RAN4   **Proposal 1: Suggest to agree on option 1 on latency requirement for data collection or inference.**  **Issue 1-7: Tests post-deployment**   * Proposals   + Option 1: The post deployment testing should be based on the model monitoring framework     - Postpone the discussion to a future release, possible as a study part of Rel-19 WI   + Option 2: RAN4 should study the ways to validate performance after model updates and/or detected drift and discuss at least the following non-mutually exclusive options:     - 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.     - 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,     - At least one fallback/default Functionality/Feature that passed conformance testing must always be present in the device.   + Option 3: There is no need for post-deployment testing   Option 4: other  **Proposal 2: Further discuss the necessity of post-deployment test (whether this can be covered by performance and LCM test) for functionality-based LCM.** |
| [**R4-2400505**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400505.zip) | Apple | For testing goals, Option 1 and/or Option 2 below will be selected depending on the test  - Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.  - FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)  - FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)  - Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.  - FFS how to define a static scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)  - FFS whether and how to define non-static specific scenarios/configurations  **Observation 1: For testing goals, both options 1 and 2 are pertinent to an assessment of performance.**  **Observation 2:** **In the context of RAN4 requirements, it is essential that generalization pertains to the capability of AI functionality to deliver a minimum expected performance across all anticipated scenarios, irrespective of the training process or the model being employed.**  **Observation 3: On device fine tuning retraining could be beneficial to model delivery/transfer to reduce overhead/latency**  **Observation 4: Having a different AI/ML model for each different Scenario/configuration could increase the UE complexity and storage requirements as well as the overhead of delivery/transfers and the associated overhead/latency.**  **Proposal 1**: **RAN4 should study the specification of reference AI/ML models for defining performance requirements for 1 and 2-sided models**  **Proposal 2: RAN4 should consider option 3 (fully specified) and option 4 (partially specified) for defining a reference encoder/decoder. Regarding option 4, RAN4 should investigate the specifications of the relevant AI model parameters. Additionally, it should specify a training dataset to ensure alignment in performance results**.  **Proposal 3 (related to Option 1): RAN4 to further discuss if both Option 1 and Option 2 should be considered for RAN4 testing purposes and to explore the primary contexts in which they are applicable and the differences between them or discuss if it would be more applicable to combine these two options with another composite testing goal that encompasses both requirements in Option 1 and 2**  **Proposal 4: A use-case and scenario dependent deterioration performance margin shall be specified for assessing if significant degradation has occurred.**  Proposal 5 RAN4 should discuss if UE can autonomously perform LCM to switch or update/finetune models during generalization tests and if this procedure is transparent  **Proposal 6:** **RAN4 to discuss the practicality of formulating a framework that facilitates on-device fine-tuning. The focus will be on exploring the feasibility of creating a dynamic and site-specific approach to online training and fine-tuning.**  **Proposal 7 RAN4 should investigate the options for enhancing the generalizability of AI/ML models by providing the appropriate assistance/side information and discuss the feasibility of training with diverse datasets**  **Proposal 8 RAN4 should clarify/agree that the side conditions of the testing procedures should remain the same for legacy and AI/ML methods.**  **Proposal 9: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case**  **Proposal 10: RAN4 shall consider the latency requirements for model monitoring input data as well as the establishment of tolerance margin requirements for the specified KPIs for model monitoring per use case**  **Proposal 11: FFS on how to perform cell level BM performance monitoring when the AI/ML model resides at NW** |
| [**R4-2400560**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400560.zip) | Qualcomm, Inc. | **Proposal 1: We propose the following principles to design the generalization tests to exclude signaling based LCM procedures and performance monitoring:**   * **Don’t include network side configuration changes, e.g., beamforming codebook and antenna configuration** * **Consider reasonable environment change within a cell site, e.g., propagation condition changes**   **Proposal 2: Agree the following principle for data collection requirement: 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.**   **Proposal 3: 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).** |
| [**R4-2401044**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401044.zip) | CMCC | ***Proposal 1: for UE-side models and/or UE-part of two-sided models, it is proposed that the scenarios and/or configurations used for generalization can be decided based on the supported configuration reported by UE.***  ***Proposal 2: for generalization, it is proposed to take the requirements for inference as the minimum level performance for generlazation.*** |
| [**R4-2401566**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401566.zip) | NTT DOCOMO, INC. | **Observation 1: Performance requirements will be specified case by case with following principle:**   * **Option 1: Whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way** * **Option 2: Whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration**   + **The legacy performance is used as baseline for existing use cases/procedures/functionalities/measurements that are to be enhanced by AI/ML based methods**   + **New or enhanced performance requirements/tests could be considered for the use cases/procedures/functionalities/measurements that are to be enhanced by AI/ML based methods without the existing legacy performance**   **Proposal 1: The simplest way to define “proper way” is that RAN4 specifies test AI/ML model and performance requirements by using the model in the specification for each test case. If DUT can select the test AI/ML model and achieve performance requirements for each case, it can be said “conducted in a proper way”.**  **Proposal 2: For option 2, a static scenario/configuration for existing use cases/procedures/functionalities/measurements should be same as existing scenario/configuration. The legacy performance should be same as existing performance requirements. On the other hand for the cases without existing legacy performance, it should be newly defined.**  **Observation 2: Generalization requirements will be specified based on following principle:**   * **Verify whether the minimum level of performance of AI/ML functionality/model can be achieved/maintain under various scenarios and/or configurations, while the performance won’t be significantly degraded in other scenarios and/or configurations.**   + **With the understanding that the propagation conditions could be covered in the various scenarios and/or configurations following the RAN1/RAN2 procedures** * **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)**   **Proposal 3: Two solutions can be considered for generalization validation.**   1. **Adding random factor (e.g. normal distribution variable) on the baseline test.**  * **Variance value has to be carefully designed**  1. **define only one additional worse test case for each use-case and fix the performance requirements as same as that of without condition change**  * **Degradation value for test condition has to be carefully designed**   **Observation 3: Performance monitoring and LCM requirements will be specified based on following principle:**   * **RAN4 to investigate how to define performance requirements/tests for the following candidate procedures:**   + **model/functionality monitoring**   + **model/functionality selection**   + **model/functionality activation/deactivation/switching/fallback** * **Define delay requirements based on multiple delay components**   **Proposal 4: LCM requirements should be defined based on procedures defined by other WGs as same as legacy functionality. However, performance monitoring metric should be defined by RAN4.**  **Proposal 5: If the model/functionality is stored in somewhere and it can be referred by TE vendor, one possible solution for post-deployment validation is that TE refers updated model/functionality and validates performance. However, it should be considered that how to treat the condition difference and how often it should be done.** |
| [**R4-2401609**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401609.zip) | vivo | ***Proposal 1: RAN4 to wait for more progress from RAN1/2 before defining core requirement for LCM procedures, e.g. functionality and model (if justified) selection, activation, deactivation, switching, fallback***  ***Observation 1: For performance monitoring, whether or not the metrics are same as inference is still an open issue under discussion.***  ***Proposal 2: RAN4 to wait more progress from RAN1/2 to consider how to define requirement for inference and performance monitoring procedures.***  ***Proposal 3: Whether and how to test/verify the performance monitoring procedure needs to be studied by RAN4.***  ***Proposal 4: Accuracy requirements for input data collection need to be considered case by case.***  ***Proposal 5:*** ***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 6: The feasibility of generalization test by using dynamic scenarios/configurations could be further studied in Rel-19 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)***   ***Observation 2: Post deployment performance may be verified by model monitoring.***  ***Proposal 7: RAN4 to further study on whether and how to test the post-deployment AI/ML feature for ensuring the***  ***performance in Rel-19 stage.***  ***Observation 3: “Ground truth” in RAN4 need to be discussed on a use case by use case basis.***  ***Proposal 8: RAN4 to study whether and how to test AI/ML feature under fading channel propagation conditions case by case.*** |
| [**R4-2401684**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401684.zip) | Huawei,HiSilicon | ***Proposal 1***: If new measurements related to performance evaluation is specified, study the feasibility, testability and necessity of defining the relevant latency requirement.  ***Observation 1***: Latency of decision-making at UE is not testable.  ***Observation 2***: Functionality/model selection and activation share similar latency/interruption if any.  ***Observation 3***: Functionality/model switching and fallback share similar latency/interruption if any.  ***Proposal 2***: Latency/interruption requirement is not applicable for functionality/model deactivation.  ***Observation 4***: Legacy requirements for existing use in RAN4 may not be applicable when defining AI/ML performance requirements, if the effect of operations from the opposite side is not eliminated or not well controlled.  ***Proposal 3***: Take functionality-based LCM as the starting point for RAN4 discussion.  ***Observation 5***: Identified scenarios and/or configurations can be initially interpreted as the scenarios and/or configurations that UE reports by capability signaling.  ***Observation 6***: A large range of various UE capabilities may be involved, which is problematic for RAN4 to identify a typical configuration/scenario for specifying the test cases.  ***Proposal 4***: RAN4 will discuss how to specify the identified scenarios and/or configurations per use case in future release, if other WGs can specify the granularity and the capability signaling.  ***Proposal 5***: According to TR 38.843, the identified scenarios and/or configurations can initially be interpreted as the scenarios and/or configurations that UE report by capability signaling.  ***Observation 7***: If legacy test metrics are not valid/testable when defining AI/ML-specific requirements, legacy performance requirements for non-AI cannot be reused.  ***Proposal 6***: RAN4 will study the minimum level performance, per use case, for identified scenarios and/or configurations (if specified).  ***Proposal 7***: Other scenarios and/or configurations are interpreted as the scenarios and/or configurations that are not reported by UE capability for an AI/ML-specific (enhanced) feature.  ***Observation 8***: There is no need to introduce AI/ML-related requirements in the other scenarios and/or configurations.  ***Proposal 8***: Performance for other scenarios and/or configurations can be ensured by RAN4 legacy test.  ***Observation 9***: There is no benefit identified by introducing channel condition changes during test.  ***Proposal 9***: Non-static scenarios/conditions and propagation conditions are precluded for defining RAN4 test. |
| [**R4-2401814**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401814.zip) | OPPO | **Proposal 1: For both CSI compression and CSI prediction, existed RAN4 test examples for “reporting of PMI” can be reused or serve as a reference. For BM and positioning, we need to consider whether/how to establish new requirements.**  **Proposal 2:** **Stability of the performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes, includes: (1) obtaining a consistent monitoring result by considering multiple evaluating samples within an evaluation window, (2) assessing whether model monitoring should be handled at the UE level or the cell level.**  **Proposal 3: In R19 RAN4 tests, we should focus on verifying whether the solution under test can work in given scenarios or conditions, and whether it can meet the requirements of RAN4 tests within these scenarios and conditions.**  **Proposal 4: In R19 RAN4 tests, static test scenarios and configurations should be considered first. After having feasible testing cases for static configurations, then further consider whether to introduce non-static testing scenarios and configurations.**  **Proposal 5: Regarding the AI/ML capabilities, following aspects should be considered**  **- Definition of basic AI/ML capability and corresponding testing requirements**  **- Definition of different AI/ML capability levels and corresponding different testing requirements**  **- Dynamic AI/ML capabilities**  **Proposal 6: FFS how to define the Post-deployment validation, what it means to achieve Post-deployment validation, and possible test methods** |
| [**R4-2402388**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402388.zip) | Samsung | *Testing goals for testing framework/procedure*  **Proposal 1: Provide the following text proposal to Option 1 of testing goal:**   * Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.   + ~~FFS~~ how to define the specific AI/ML model is provided as use-case specific manner (e.g., a model captured in RAN4 spec as baseline, or a reference model structure agreed for performance alignment)   + ~~FFS~~ how to define that the model is properly conducted is provided as use-case specific manner (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)   *Relation to legacy requirements*  **Observation 1: Based on existing agreement, for the case with the existing legacy performance, the AI/ML enhanced performance shall be defined by using legacy requirement as baseline.**  **Proposal 2: No further study or TR refinement is needed for the general aspect of “relation to legacy requirements”, and detailed requirement can be discussed based on specific use case.**  *Generalization aspects*  **Proposal 3: Further understanding on generalization goals is provided/highlighted:**   * details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality   + - Different scenarios used for generalization test are served as the additional conditions for the AI/ML model training but are not a part of UE capability for the AI/ML-enabled feature/FG     - Configurations used for generalization test shall all be associated with UE capability of an AI/ML-enabled Feature/FG   **Proposal 4: No further study or TR refinement as general principle for testability is needed for the following two aspects for generalization test:**   * **what the minimum level performance for each identified scenario and/or configuration is** * **what the significant degradation for other scenarios and/or configurations is**   *Propagation conditions for testing*  **Proposal 5: Provide the following text proposal to field dataset for the propagation conditions for testing:**   * Field dataset (data collected directly from field measurements)   + Feasibility of using field dataset can only be confirmed if the dataset is well defined and open to DUT vendors   *Static/non-static scenarios/conditions for testing*  **Proposal 6: For how to define a static scenario/configuration:**  **(1) the scenario/configuration should be determined by “the specific configuration/conditions” associated with the relevant “UE capability of an AI/ML-enabled Feature/FG” under testing;**  **(2) the static scenario/configuration determined in (1) shall be maintain unchanged during the test.**  *UE processing capability and limitations*  **Proposal 7: Further study on AI/ML relevant UE processing capability and limitations by considering:**  **1) enabling two or more AI/ML functionalities for concurrent operation or**  **2) enabling the combination of certain AI models for single or multi-functionalities.**  *Post-deployment validation*  **Proposal 8: Before any deployment to the UE is performed, UE vendor shall already perform test(s) (in-house development test and/or conformance test) on the to-be-deployed AI/ML model, therefore post-deployment validation (before activation) is not needed.** |
| [**R4-2402412**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402412.zip) | Ericsson | [Observation 1 Existing RF and RRM requirements are expected to be achieved in all circumstances.](#_Toc159254993)  [Observation 2 Existing performance requirements are defined under certain conditions, but when operating in the real-world similar performance is to be expected in other, similar conditions.](#_Toc159254994)  [Observation 3 RAN4 does not need to directly consider generalization. RAN4 needs to ensure that minimum performance is observed across all applicable deployment conditions](#_Toc159254995)  [Observation 4 It is likely that a robust monitoring procedure is needed to complement conformance testing to manage performance in all conditions.](#_Toc159254996)  [Observation 5 Care should be taken that non-dynamic testing captures properly any model switching behaviors.](#_Toc159254997)  [Observation 6 RAN4 may need to consider whether performance of different AI functionalities is independent or not.](#_Toc159254998)  Based on the discussion in the previous sections we propose the following:  [Proposal 1 RAN4 to discuss whether using legacy as baseline implies achieving similar performance to a non-AI that meets requirements over all expected conditions, not just the test configuration.](#_Toc159254999)  [Proposal 2 RAN4 to discuss (probably on a use-case specific basis) how to define the range of conditions over which baseline performance should be achieved and how much variation in performance there might be over different conditions.](#_Toc159255000)  [Proposal 3 RAN4 to consider whether it would be useful to set a “minimum SNR vs throughput” that would be applicable in any relevant channel condition (a bit similar to core RF/RRM requirements).](#_Toc159255001)  [Proposal 4 For each use case, there is a need to study how the AI model behaviour changes with different scenarios (for example, whether the change is smooth)](#_Toc159255002)  [Proposal 5 For each use case, RAN4 needs to study whether using synthetic channels for test data will reliably test models trained on real data.](#_Toc159255003)  [Proposal 6 RAN4 should study accuracy requirements on monitoring reports if/as defined by RAN1/2. Feedback to RAN1/2 may be needed on which metrics could be feasible to set requirements on and which not.](#_Toc159255004)  [Proposal 7 RAN4 should consider further how a test of a model quality report could work, since different model quality levels would need to be forced.](#_Toc159255005) |
| [**R4-2402439**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402439.zip) | Intel Corporation | **Proposal #1: UEs supporting AI/ML features shall be mandated to meet the existing legacy performance requirements with configured/enabled AI/ML functionality for all existing legacy test cases.**  **Proposal #2: The minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature shall be tested for a static and/or non-static scenario and configuration:**   * + - **Static scenario/configuration term means that at least channel model type and SNR settings are fixed and do not change over the test, while specific channel realizations may be dynamic.**     - **Static scenarios/configurations can be applicable to all use cases.**     - **Non-static scenarios/configuration can be further considered in application to CSI and beam management temporal prediction use cases. The details of models are FFS.**   **Proposal #3: Adopt the following framework for post-deployment model (feature/functionality) verification at least for the case when model updates or changes are non-transparent to the network:**   * + - **At least some default AI/ML model (feature or functionality) needs to pass conformance testing and be present in the device.**     - **Any changes or updates to the ML-enabled functionality or feature may be tested by the device vendor against RAN4 requirements before any deployment to the UE is performed.**     - **The post deployment testing should be based on the model monitoring framework.**     - **The information on whether AI/ML model update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**   **Proposal #4: The complexity of AI/ML models and the capabilities of UE should be taken into account when defining minimum requirements with no need for additional studies during study stage.**  **Proposal #5: Concurrent AI/ML feature testing should be handled with the 2nd priority in favor of individual feature or model verification.** |
| [**R4-2402565**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402565.zip) | Nokia, Nokia Shanghai Bell | The following Observations and Proposals were made:  **On AI/ML Requirements/Testing Framework:**  **Proposal 1: RAN4 to consider performance requirements for Performance monitoring (in addition to Core), depending on RAN1/RAN2 design of corresponding procedures.**    *Figure 1: Illustration of RAN4 AI/ML Requirements/Testing Framework.*  **Proposal 2: RAN4 to include a diagram like in Figure 1 into the TR 38.843 to illustrate RAN4 AI/ML requirements/testing framework.**  **Observation 1:** The definitions of AI/ML Model testing and validation introduced in TR 38.843 are not aligned with RAN4. In particular, RAN4 AI/ML-based feature testing cannot be the subprocess of training.  **Proposal 3: Add a note in the term definitions (Clause 3.1 of TS 38.843) of AI/ML model testing and AI/ML model validation that they are not applicable in RAN4 context.**  A diagram of a software system  Description automatically generated  *Figure 3: Block diagram of a ML-enabled feature/functionality testing.*  **Observation 2:** Evaluation methodologies that are considered to isolated AI/ML models (e.g. in RAN1) cannot be directly applied in RAN4 for formulation of requirements and test cases.  **Proposal 4: RAN4 to adopt Figure 3 demonstrating the testing principles of AI/ML-enabled feature/functionality as AI/ML-enabled system in the TR 38.843.**  **Proposal 5: RAN4 configurations/parameters, requirements and tests should be defined on the level of ML-enabled Functionality/Feature, i.e., model- specific requirements and tests shall be precluded.**  **Proposal 6: RAN4 to consider introducing the definition of Requirement/reference AI/ML model – the AI/ML model that is used by a particular company for the alignment/derivation of RAN4 performance requirements.**  **On generalization:**  **Observation 3:** The possibility to distinguish Identified and other Conditions for generalization depends on the way how UE capabilities will be defined for AI/ML-enabled features and what (additional) information can be shared in between the UE and the NW.  **Proposal 7: RAN4 to address minimum level of performance and potential level of degradation for generalization requirements in use-case specific discussions.**  **On Core signaling requirements for LCM:**  **Observation 4:** If an LCM action is required and it is not taken in a timely manner, the performance degradation for AI/ML enabled use cases may be degraded to undesirable levels.  **Proposal 8: RAN4 LCM core requirements on AI/ML functionality managements should be addressed with higher priority to guarantee latency of UE behavior/response when receiving functionality activation, deactivation, fallback/switching actions (RRC/MAC-CE/DCI signalling).**  **Observation 5:** Model training data collection mechanisms are strongly dependent on the progress in RAN1 and RAN2.  **Observation 6:** Design of model identification mechanism still need more clarity from RAN1 and RAN2.  **Proposal 9: RAN4 to wait for further progress in RAN1 and RAN2 before discussing any testing requirements and procedures which involve AI/ML model identification and AI/ML model training data collection.**  **Proposal 10: RAN4 not to discuss testing requirements for model transfer for the UE-sided beam management and positioning accuracy enhancement use cases. RAN4 to wait for further progress in RAN1 and RAN2 before discussing testing requirements for UE-part model transfer/update in the 2-sided CSI feedback enhancement use cases.**  **On Model updates and post-deployment verification:**  **Proposal 11: RAN4 should study a framework to enable post deployment verification and/or tests for model updates 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. **Rely on Performance monitoring mechanism, but assume that at least one fallback/default Functionality/Feature that passed conformance testing shall be present in the device.**   **On UE processing capabilities requirements**  **Proposal 12:** **RAN4 should test the mutual impact of several simultaneously supported and active ML-enabled Functionalities/Features to ensure the absence of performance degradation.**  **On testing diagrams:**  **Proposal 13: RAN4 to agree and clarify in the TS that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 [3] is applicable only for the testing UE-sided model -enabled use cases.**  **Proposal 14: Proposal 13: RAN4 to agree that the description of the reference block diagram in Figure 7.3.2.3-1 in TR38.843 requires clarifications for at least the following items:**  **• The role and meaning of the ‘LCM’ blocks in the TE and DUT**  **• The role and meaning of the ‘AI/ML functions’ block in the TE**  **• The role and meaning of the ‘Inference’ block in the DUT**  **• The role and meaning of the ‘Verification’ block in the TE**  **• The signaling/messages assumed on the physical links between the TE/gNB and DUT/UE**  **A diagram of a computer system  Description automatically generated** |

## Open issues summary

The open issues were grouped in the following sub-topics for further discussion based on the input contributions. Considering that this is the last meeting of the SI, some of the topics are urgent to finalize the TR of the SI and be able to complete. These topics are marked as 1st priority. The 2nd priority items should be discussed if time allows in order to enable further progress and improve the understanding of the group related to this rather complex study.

1st prirotity:

1. Generalization
2. Post deployment testing
3. Latency requirements
4. Data collection principles
5. On-device training/fine-tuning
6. Combination of capabilities
7. Performance in synthetic channels vs. real data
8. Updates to testing goals
9. Testing diagram update/clarification
10. AI/ML Requirements/testing framework diagram

### Sub-topic 1-1

*Generalization*

Several companies brought up the issue of generalization. Some of the proposals to be discussed are listed below.

**Issue 1-1: Generalization update**

* Proposals
  + Option 1 – Proposal 5 from R4-2400505:
* Proposal 5 RAN4 should discuss if UE can autonomously perform LCM to switch or update/finetune models during generalization tests and if this procedure is transparent
  + Option 2 – Proposal 1 from R4-2400560

**Proposal 1: We propose the following principles to design the generalization tests to exclude signaling based LCM procedures and performance monitoring:**

* **Don’t include network side configuration changes, e.g., beamforming codebook and antenna configuration**
* **Consider reasonable environment change within a cell site, e.g., propagation condition changes**
  + Option 3: Proposal 1 from R4-2401044
* ***Proposal 1: for UE-side models and/or UE-part of two-sided models, it is proposed that the scenarios and/or configurations used for generalization can be decided based on the supported configuration reported by UE.***
  + Option 4: Proposal 3 from R4-2401566:

**Proposal 3: Two solutions can be considered for generalization validation.**

1. **Adding random factor (e.g. normal distribution variable) on the baseline test.**

* **Variance value has to be carefully designed**

1. **define only one additional worse test case for each use-case and fix the performance requirements as same as that of without condition change**

* **Degradation value for test condition has to be carefully designed**
  + Option 5: Proposal 6 from R4-2401609

***Proposal 6: The feasibility of generalization test by using dynamic scenarios/configurations could be further studied in Rel-19 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)***
  + Option 6: Proposals 3/4 in R4-2402388

**Proposal 3: Further understanding on generalization goals is provided/highlighted:**

* details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality
  + - Different scenarios used for generalization test are served as the additional conditions for the AI/ML model training but are not a part of UE capability for the AI/ML-enabled feature/FG
    - Configurations used for generalization test shall all be associated with UE capability of an AI/ML-enabled Feature/FG

**Proposal 4: No further study or TR refinement as general principle for testability is needed for the following two aspects for generalization test:**

* **what the minimum level performance for each identified scenario and/or configuration is**
* **what the significant degradation for other scenarios and/or configurations is**
  + Option 7: No need for any agreement, discuss on a case by case basis for each feature or test (see also Proposal 7 in R4-2402565
* Recommended WF

To be discussed

### Sub-topic 1-2

*Post deployment testing*

Post deployment testing/validation was discussed in RAN4#109 with the following options:

* + Option 1: The post deployment testing should be based on the model monitoring framework
    - Postpone the discussion to a future release, possible as a study part of Rel-19 WI
  + Option 2: RAN4 should study the ways to validate performance after model updates and/or detected drift and discuss at least the following non-mutually exclusive options:
    - 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.
    - 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,
    - At least one fallback/default Functionality/Feature that passed conformance testing must always be present in the device.
  + Option 3: There is no need for post-deployment testing
  + Option 4: other

The issue should be further discussed.

**Issue 1-2: Post deployment handling**

* Proposals
  + Option 1: Proposal 5 in R4-2401566

**Proposal 5: If the model/functionality is stored in somewhere and it can be referred by TE vendor, one possible solution for post-deployment validation is that TE refers updated model/functionality and validates performance. However, it should be considered that how to treat the condition difference and how often it should be done.**

* + Option 2: Post deployment performance may be verified by model monitoring, postpone the discussion until LCM framework becomes clear.
  + Option 3: Proposal 8 in R4-2302388
* **Proposal 8: Before any deployment to the UE is performed, UE vendor shall already perform test(s) (in-house development test and/or conformance test) on the to-be-deployed AI/ML model, therefore post-deployment validation (before activation) is not needed.** 
  + Option 4: Proposal#3 in R4-2402439:

**Proposal #3: Adopt the following framework for post-deployment model (feature/functionality) verification at least for the case when model updates or changes are non-transparent to the network:**

* + - **At least some default AI/ML model (feature or functionality) needs to pass conformance testing and be present in the device.**
    - **Any changes or updates to the ML-enabled functionality or feature may be tested by the device vendor against RAN4 requirements before any deployment to the UE is performed.**
    - **The post deployment testing should be based on the model monitoring framework.**
    - **The information on whether AI/ML model update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**
  + Option 5: Proposal 11 in R4-2402565:

**Proposal 11: RAN4 should study a framework to enable post deployment verification and/or tests for model updates 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. **Rely on Performance monitoring mechanism, but assume that at least one fallback/default Functionality/Feature that passed conformance testing shall be present in the device.**
   * Option 6: No need for any post deployment

* Recommended WF
  + To be discussed

### Sub-topic 1-3

*Latency for data collection or inference*

Latency requirements were discussed in RAN4#109 and several options were debated. Some proposals on handling latency were also input into this meeting. The options below are not exclusive, they should all be discussed to see if any agreements can be reached.

**Issue 1-3: Latency requirements**

* Proposals
  + Option 1: Latency requirements of data collection for model inference and monitoring could be considered per use case , subject to output from RAN1/2.
    - This includes new measurements or reports to be defined
  + Option 2: Latency of decision making at UE is not testable
  + Option 3: Latency/interruption requirement is not applicable for functionality/model deactivation.
  + Option 4: Others
* Recommended WF
  + Agree Option 1, discuss whether any other agreements can be made

### Sub-topic 1-4

*Data collection principles*

Some companies have raised the issue of how to handle data collection even though this could be too early.

**Issue 1-4: Data collection**

* Proposals
  + Option 1: Agree the following principle for data collection requirement: 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.
  + Option 2: Accuracy requirements for input data collection need to be considered case by case.
  + Option 3: RAN4 to wait for further progress in RAN1 and RAN2 before discussing any testing requirements and procedures which involve AI/ML model identification and AI/ML model training data collection.
  + Option 4: Other proposals
* Recommended WF

To be discussed

### Sub-topic 1-5

*On device training/fine-tuning*

The issues of handling of on device training or fine-tuning of AI/ML models was raised.

**Issue 1-5: On device training/fine-tuning**

* Proposals
  + Option 1: RAN4 to discuss the practicality of formulating a framework that facilitates on-device fine-tuning. The focus will be on exploring the feasibility of creating a dynamic and site-specific approach to online training and fine-tuning.
  + Option 2: One device training will lead to over-fitting, this should not be enabled
  + Option 3: Framework is not needed
  + Option 4: Other proposals
* Recommended WF

To be discussed

### Sub-topic 1-6

*Handling of combination of features/capabilities*

**Issue 1-6: Combinations of features/capabilities**

* Proposals
  + Option 1: Further study on AI/ML relevant UE processing capability and limitations by considering:

1) enabling two or more AI/ML functionalities for concurrent operation or

2) enabling the combination of certain AI models for single or multi-functionalities.

* + Option 2: Concurrent AI/ML feature testing should be handled with the 2nd priority in favor of individual feature or model verification.
  + Option 3: RAN4 should test the mutual impact of several simultaneously supported and active ML-enabled Functionalities/Features to ensure the absence of performance degradation.
  + Option 4: The number of combinations can be very high, combination of features should not be considered
  + Option 5: other
* Recommended WF

To be discussed

### Sub-topic 1-7

*Test data handling*

The issue of whether synthetic channels would be adequate for performance assessment/testing was brought up and it would be useful to discuss

**Issue 1-7: Test data handling**

* Proposals
  + Option 1: For each use case, RAN4 needs to study whether using synthetic channels for test data will reliably test models trained on real data.
  + Option 2: Synthetic channels (RAN4 channels) should be enough for testing
  + Option 3: RAN4 should discuss whether/how field data can be used for testing
  + Option 4: Other
* Recommended WF

To be discussed

### Sub-topic 1-8

Testing goals update

The testing goals were already captured in the TR in RAN4#110, some updates are proposed as follows

**Issue 1-8: Testing goals updated**

* Proposals
  + Option 1: Updated Option 1 of testing goals as follows (R4-2402388):
* Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.
  + ~~FFS~~ how to define the specific AI/ML model is provided as use-case specific manner (e.g., a model captured in RAN4 spec as baseline, or a reference model structure agreed for performance alignment)
  + ~~FFS~~ how to define that the model is properly conducted is provided as use-case specific manner (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)
  + Option 2: no updated is needed
  + Option 3: others, please provide some concrete proposals
* Recommended WF

To be discussed

### Sub-topic 1-9

Testing diagram update/clarification

**Issue 1-9: Testing goals updated**

* Proposals
  + Option 1:
    - Proposal 13: RAN4 to agree and clarify in the TS that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 [3] is applicable only for the testing UE-sided model -enabled use cases.
    - Proposal 14: Proposal 13: RAN4 to agree that the description of the reference block diagram in Figure 7.3.2.3-1 in TR38.843 requires clarifications for at least the following items:
      * The role and meaning of the ‘LCM’ blocks in the TE and DUT
      * The role and meaning of the ‘AI/ML functions’ block in the TE
      * The role and meaning of the ‘Inference’ block in the DUT
      * The role and meaning of the ‘Verification’ block in the TE
      * The signaling/messages assumed on the physical links between the TE/gNB and DUT/UE

**A diagram of a computer system

Description automatically generated**

* + Option 2: Other updates
  + Option 3: no update needed
* Recommended WF

To be discussed

### Sub-topic 1-10

AI/ML Requirements/testing framework diagram

A diagram summarizing the requirements and testing framework is proposed.

**Issue 1-10: Testing goals updated**

* Proposals
  + Option 1: **RAN4 to include a diagram like in Figure 1 into the TR 38.843 to illustrate RAN4 AI/ML requirements/testing framework.**



* *Figure 1: Illustration of RAN4 AI/ML Requirements/Testing Framework.*
  + Option 2: Changes are needed to the diagram
  + Option 3: Such diagram is not needed
* Recommended WF

To be discussed

# Topic #2: Testability and interoperability issues for beam management

This section contains the sub-topics regarding specific issues for beam management.

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2400091**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400091.zip) | CATT | **Proposal 1: The BM use case tests for performance, LCM and generalization should be defined separately.**  **Proposal 2: One general performance** **test procedure is possible for all alternatives for two BM sub-use cases, i.e., no need to define different test procedures for different BM sub-use cases.**  **Proposal 3: During the tests, the data for inference can be obtained by measuring the beams in Set B transmitted by TE if DUT is UE, or are directly provided by TE if DUT is gNB. The specific data type depends on the input data type required by the deployed models, e.g., L1-RSRP.**  **Proposal 4: For monitoring or verification purpose, the following two alternatives can be considered where two types of data are required in each alternative:**  **- Inferred data and ground truth, e.g., inferred best beam ID and the actual best beam ID in Set A, or**  **- An assessment value and a threshold, e.g., successful rate of correct prediction and a pre-defined threshold.** |
| [**R4-2400134**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400134.zip) | CAICT | **Proposal 1: Legacy measurement performance requirements of L1-RSRP could be considered as starting point for data collection of BM. New requirements could be considered if necessary, according to RAN1 conclusion.**  **Proposal 2: Suggest to discuss which entity can provide the reference value for evaluating inference performance and candidate solutions to obtain the value.**  **Proposal 3: Suggest to discuss feasible monitoring metrics in practical networks to verify that the performance of AI/ML model could be properly monitored.** |
| [**R4-2400506**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400506.zip) | Apple | For metrics for beam management requirements/tests, the following test metrics are identified and could be considered  - Option 1: RSRP accuracy  - Option 2: Beam prediction accuracy  - Top-1 (%) : the percentage of "the Top-1 strongest beam is Top-1 predicted beam"  - Top-K/1 (%) : the percentage of "the Top-1 strongest beam is one of the Top-K predicted beams"  - Top-1/K (%) : the percentage of "the Top-1 predicted beam is one of the Top-K strongest beams"  - Option 3: The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB,  - Related measurement accuracy can be considered to determine x  - Option 4: combinations of above options  **Observation 1: For testing purposes and defining requirements for BM, a particular UE can pass a test under one test metric but fail the other Options/test metrics.**  **Observation 2: Ensuring identical beam IDs between legacy/genie system and AI/ML models for RSRP comparisons, as well as maintaining an average RSRP accuracy across multiple beams, poses a challenge for the Option 3 test metric. (that is, it fails to test for those)**  **Observation 3: For option 2 test metric, using the beam ID prediction accuracy as the KPI for validating the test could lead to testing issues since RSRP accuracy couldn’t be guaranteed or tested**  **Observation 4: RAN4 needs to capture a different test metric or a combination of Options 1,2,3**  **Observation 5: The source of training data for beam management will play a crucial role in AI/ML BM performance and in the generalization performance in real deployment**  **Observation 6:** **For training data based on real measurements, the** **quality of training data depends on RF impairments, and other noise sources. There is tradeoff between training data quality and generalization performance.**  **Observation 7:** **To guarantee that the UE operates within acceptable margins, it's essential to subject it to various radio conditions and configurations for testing and generalization validation in RAN4**  **Proposal 1: For testing purposes and defining requirements for BM, RAN4 should clarify if RSRP accuracy pertains to the scenario where the top predicted AI/ML Beam ID is the same with legacy’s/genie**  **Proposal 2: There are issues with Option 3 test metric for beam management requirements/tests. RAN4 to revise option 3 to address the issues discussed or propose a different test metric.**  **Proposal 4: For testing purposes and defining requirements for BM, RAN4 to specify a new KPI that is a combination of Options 1,2,3**  **Proposal 5: RAN4 to investigate the source of training data for BM by taking into consideration the advantages and disadvanges of all the considered options**  **Proposal 6:**  **For training data based on real measurements, for BM-Case1 and BM-Case2, RAN4 should study the impact of legacy L1-RSRP measurement accuracy requirements (accuracy of training data) as well as the quality of those data on the performance and generalization of AI/ML based BM**  **Proposal 7: Reference radio conditions and configuration/parameter settings need to be specified for the Beam Management (BM) use case. In addition to the reference conditions and configurations, various other scenarios shall be agreed upon for testing generalization**  **Proposal 8: To improve the generalizability of the AI/ML model and reduce the complexity and requirements for model delivery/transfer, RAN4 should explore integrating side/assistance information within the AI/ML model** |
| [**R4-2400561**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400561.zip) | Qualcomm, Inc. | **Observation 1: An effective test case for beam prediction use case BM-Case1 requires the following:**   * **Sufficient randomness and variation in time and spatial domain of L1-RSRP has to be emulated in the test** * **Enough number of Tx beams in Set B and Set A from different (peak direction) AoAs with different beam patterns on UE’s sphere coverage has to be emulated in the test**   **Observation 2: RAN4 L1-RSRP and other measurement test configurations support only:**   * **Deterministic power configuration on each AoA** * **Tx beams up to 2 AoAs**   **Observation 3: The following conditions contribute to randomness and variation in time and spatial domain of L1-RSRP**   * **Propagation conditions as a function of (1) AoD of the Tx beam (2) AoA of the Rx beam (3) fading condition** * **Tx beamforming gain on the AoDs in the propagation conditions** * **UE movement**   **Proposal 1: Based on the agreements in TR 38.843 on beam prediction quoted above, we propose to study the test feasibility for beam prediction from the following perspectives:**   * **How to model the randomness and variation in time and spatial domain of L1-RSRP by the propagation conditions and the Tx beamforming gain on the AoDs considered in the propagation conditions** * **The verification procedures to confirm that the power level probability distributions on different AoAs aligns with the probability distributions derived from the configured propagation conditions and Tx beamforming gain** * **The number of beams and their coverage in spatial domain in Set B and Set A required to test beam prediction performance**   **We propose the following setup as the starting point for the feasibility studies**   * **Consider CDL channel to model the randomness in the propagation conditions so that AoD, AoA and fading can be captured** * **Consider the antenna configuration from TR 38.901 clause 7.3 to design the beam peak directions, derive the corresponding beam patterns and determine the Tx beamforming gain on the concerned AoDs in the propagation conditions** * **Consider to define TE verification procedures to ensure that the power level probability distributions on different AoAs aligns with the probability distributions derived from the configured propagation conditions and Tx beamforming gain** * **Consider emulating UE rotation in the beam prediction test to model the AoA changes from UE perspective.**   **Proposal 2: The following issues should be considered when defining the beam prediction accuracy requirements**   * **Consistency between training and testing data (from the perspectives of beams in Set B and Set A, physical characteristics of gNB antenna etc.) should be guaranteed by signaling conveyed to UE.** * **The impact of size and composition of Set B and Set A on accuracy requirement.** |
| [**R4-2401046**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401046.zip) | CMCC | ***Proposal 1: for beam management, it is proposed to take RSRP accuracy and beam prediction accuracy as metrics for requirements/tests.***  ***Proposal 2: when specify performance reqirements for AI/ML based beam management, it is proposed to discuss whether and how to consider the impact due to different assumption, e.g. quantization granularity of L1-RSRP, measurement error, etc.*** |
| [**R4-2401171**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401171.zip) | xiaomi | For testing goals, Option 1 and/or Option 2 below will be selected depending on the test   * Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.   + FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)   + FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs) * Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.   + FFS how to define a static scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)   FFS whether and how to define non-static specific scenarios/configurations  **Proposal 1: Both option 1 and option 2 based test goal are needed for beam management.**  **Proposal 2: RAN4 to wait for RAN1 progress on conditions for functionality/model identification to define detail test scenario.**  **Proposal 3: Apply the dataset based on TR 38.901 for beam management test in RAN4.**  **Proposal 4: If RSRP prediction accuracy will be tested, RAN4 to further discuss how to define ground truth of L1-RSRP.**  **Proposal 5: If beam prediction accuracy will be tested, the ground truth of best TX beam index is Top-K genie-aided Tx beam, which is known by generator and is not measured Top-K Tx beam.** |
| [**R4-2401610**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401610.zip) | vivo | ***Proposal 1: RAN4 to use RSRP accuracy as the baseline KPI to evaluate AI/ML based beam management inference performance and further study whether it needs combination with Option 2/3 in WI***  ***Proposal 2: RAN4 to study if reference model is needed in BM for alignment on KPI performance***  ***Observation 1: It will bring challenges to TE for supporting tests of multiple beams in Set B (e.g., num=16,32,64)***  ***Observation 2: In the legacy L1-RSRP measurement performance requirements, the AWGN channel is assumed as the propagation condition in the test, for which the comparable ground truth values of upper and lower bound can be easily obtained. However, if the propagation condition is changed to fading channel, it would make it challenging to obtain.***  ***Proposal 3: RAN4 to consider verify the BM performance under AWGN channel as the baseline and further study the feasibility on testing under the fading channel, e.g., how to obtain the ground truth values of upper and lower bound on range in test requirement under fading channel environment.***  ***Observation 3: For temporal DL beam prediction, how to figure out the testing method would be challenging if UE trajectory model is considered***  ***Proposal 4: Deprioritize to considered UE trajectory modeling in temporal DL beam predication test in initial phase of WI*** |
| [**R4-2401685**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401685.zip) | Huawei,HiSilicon | ***Proposal 1:*** For NW-sided beam prediction, the requirement of RSRP measurement reporting reuses RAN4 legacy.  For metrics for beam management requirements/tests, the following test metrics are identified and could be considered   * Option 1: RSRP accuracy * Option 2: Beam prediction accuracy   + Top-1 (%): the percentage of “the Top-1 strongest beam is Top-1 predicted beam”   + Top-K/1 (%): the percentage of “the Top-1 strongest beam is one of the Top-K predicted beams”   + Top-1/K (%): the percentage of “the Top-1 predicted beam is one of the Top-K strongest beams” * Option 3: The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB,   + Related measurement accuracy can be considered to determine x * Option 4: combinations of above options   The overhead/latency reduction should be considered for the requirements as the side condition.  ***Proposal 2:*** For UE-sided beam prediction, take Option 1 as baseline to ensure that the received RSRP is larger than a pre-defined threshold, if TE uses the predicted beam for DL transmission.  ***Observation 1:*** The test metric in Option 2 is not applicable if the beam measurement accuracy tolerance is not considered.  ***Proposal 3:*** For UE-sided beam prediction, deprioritize the test metric of beam prediction accuracy in Option 2. |
| [**R4-2401815**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401815.zip) | OPPO | **Proposal 1: Performance requirement on BM model/functionality input (e.g. beam measurement accuracy) should be supported.**  **Proposal 2: Option 1(RSRP accuracy) should be utilized to test the BM model/functionality output.**  **Proposal 3: UE can monitor the performance of AI/ML based CSI model/functionality through RSRP accuracy.**  **Proposal 4: Stability of the performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes.**  **Proposal 5: For BM testability, RAN4 need to:**  **(1) Clarify the limitation on the FR2 beams that TE can support(or TE can simultaneous support),**  **(2) Consider how to ensure that the BM model constructed on the DUT side can be matched(or approximate matched) and utilized in the testing environment on the TE side.** |
| [**R4-2401818**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401818.zip) | ZTE Corporation | **Observation 1: The UE shall report the predicted RSRP corresponding to predicted beams ID, then the TE will check whether the range of predicted RSRP of Top-K beams includes the RSRP value of strongest beam.**  **Observation 2: The UE will report the predicted beam ID, then the TE will check whether predicted Top-K beams ID includes the strongest beam.**  **Proposal 1: The RSRP accuracy of Top-K or Top-1 predicted beams shall be used as performance metrics for AI/ML based beam management.**  **Proposal 2: Top-K or Top-1 beam ID prediction accuracy shall be used as performance metrics for AI/ML based beam management.** |
| [**R4-2401920**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401920.zip) | MediaTek Inc. | ***Observation 1***: **It is possible to get the ideal L1-RSRP during the test in FR1.**  ***Observation 2***: **It is impossible to get the ideal L1-RSRP after Rx beamforming during the test in FR2.**  ***Proposal 1***: **In the test for DL Tx beam prediction, the RSRP level at TE side can be used as the reference to check the prediction accuracy or RSRP accuracy in FR1.**  ***Proposal 2***: **In the test for DL Tx beam prediction, the measured results by UE can be used as the reference to check the prediction accuracy or RSRP accuracy in FR2.**  ***Proposal 3***: **Wait for more RAN1/2 progress to discuss whether to test LCM related procedures for DL Tx beam prediction.**  ***Proposal 4***: **For DL Tx beam prediction, the testing goal should at least verify the minimum performance of the AI/ML model.**  ***Proposal 5***: **To align with the real field, the reference UE behavior should consider measurement error added at both input and output in training.**  ***Proposal 6***: **For DL Tx beam prediction, the impact of measurement error should be further evaluated, e.g.,** **further discuss how to model reasonable measurement error when using a testing dataset based on channel models in TR 38.901.** |
| [**R4-2402304**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402304.zip) | Nokia, Nokia Shanghai Bell | **Observation 1:** Option 1 of performance metrics (RSRP accuracy) may be valid for all reported beams measurements and not only for the top-1 or top-K beams measurements.  **Observation 2:** It should be noted that any error in measurement accuracy would impact the performance of AI/ML enabled BM-Case1 and BM-Case2. Furthermore, these measurement errors would potentially impact performance monitoring aspects as well, which are under discussion in RAN1.  **Proposal 1: RAN4 should prioritize option 1 (RSRP accuracy) and option 2 (Beam prediction accuracy) for performance requirement metric of AI/ML Enabled BM-Case1 and BM-Case2.**  **Observation 3:** Different Options, under discussion in RAN1 for performance monitoring aspects of BM, would potentially have different impacts in RAN4.  **Proposal 2: RAN4 to follow RAN1 progress on performance monitoring to evaluate any impacts on RAN4 for BM-Case1 and BM-Case2.**  **Proposal 3: RAN4 should further discuss the test mechanisms, test setups and their feasibility for performance metrics for AI/ML enabled BM-Case1 and BM-Case2 (i) RSRP accuracy, (ii) Beam prediction accuracy. RAN5 feedback about the appropriate test setups for these test mechanisms might also be needed.**  **Observation 4:** If an LCM action is required and it is not taken in a timely manner, the performance for AI/ML enabled BM use case may be degraded to undesirable level.  **Proposal 4: RAN4 core requirements should be considered to limit latency of LCM actions (activate/de-activate/switch/fallback to legacy) towards the DUT/UE for AI/ML enabled BM-Case1 and BM-Case2.**  **Proposal 5: RAN4 should defined identified scenarios and other scenarios for AI/ML enabled BM-Case1 and BM-Case2. Furthermore, RAN4 should also define minimum level of performance for identified scenarios and acceptable tolerance margin for other scenarios.** |
| [**R4-2402389**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402389.zip) | Samsung | *RAN4 Core Requirements for AI-BM*  The summary on the expected RAN4 requirement impact and correspondingly our proposals is provided as:   |  |  |  |  | | --- | --- | --- | --- | | UE/NW | Operations | RAN4 Issues | Samsung Proposal | | NW-sided model | Data collection | (1) RAN4 requirement for enhancement on UE measurement and report to support data collection (i.e., the contents of the collected data) | Proposal 1 (Necessary core requirement for data collection) | | (2) RAN4 requirement for supporting overhead reduction (the omission/selection of collected data, and/or the compression of collected data), if introduced. | | Inference | N/A (gNB-Implementation based) | Proposal 2 (No RAN4 impact expected) | | Perf. monitoring | N/A (gNB-Implementation based) | Proposal 3 (No RAN4 impact expected) | | LCM | (3) RAN4 requirement impact, by considering the design of signaling/mechanism(s) for LCM | Proposal 4 (FFS, depends on RAN2 introduced signaling/mechanism if any) | | UE-sided model | Data collection | N/A at least for data collection for model training (UE-Implementation based) | Proposal 5 (No RAN4 requirement for training) | | Additional assistance information | (4) Impact on RAN4 requirement to ensure consistency/association between training and inference regarding NW-side additional assistance information for inference at UE. | Proposal 6 (Consistency/association shall be guaranteed in RAN4 requirement) | | Inference | (5) RAN4 requirement for supporting model inference. | Proposal 7 (Necessary core requirement for model inference with new RAN1 mechanisms) | | Perf. monitoring | (6) RAN4 requirement for supporting performance monitoring. | Proposal 8 (RAN4 requirement on Type 1 (Option 2) performance monitoring is required.) | | LCM | (7) RAN4 requirement for supporting LCM. | Proposal 9 (RAN4 requirement for LCM performed by the gNB is required) |   where the following observation/proposals are provided for NW-sided model:  **Observation 1: From RAN1 perspective, the enhancement on UE measurement and reporting (i.e., increased number of reported RSRPs and the increased number of measured beams/CSI-RS resources) could be introduced for Rel-19 AI-BM.**  **Observation 2: From RAN1 perspective, the approaches for overhead reduction could be introduced for Rel-19 AI-BM.**  **Observation 3: For data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, L1 signaling or higher-layer signaling could be used to report the collected data.**  **Proposal 1: RAN4 shall introduce the necessary core requirement on supporting data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, by considering the new aspects including:**   1. **enhancement on UE measurement and reporting;** 2. **overhead reduction on UE reporting;** 3. **L1 or higher-layer signalling design.**   **Proposal 2: No RAN4 impact is expected for model inference for BM-Case1 and BM-Case2 with a NW-side AI/ML model.**  **Proposal 3: No RAN4 impact is expected for performance monitoring for BM-Case1 and BM-Case2 with a NW-side AI/ML model.**  **Proposal 4: RAN4 shall further study the necessary RAN4 requirements, depending on RAN2-introduced signalling/mechanism, if any.**  and the following observation/proposals are provided for UE-sided model:  **Observation 4: The potential interaction(s) between UE and NW for data collection for UE-sided model could be: (1) UE reporting of the supported/preferred DL RS configuration; (2) trigger/initiating data collection; (3) assistance information from NW to UE for data collection.**  **Proposal 5: RAN4 shall not define any requirement for data collection for UE-sided model training.**  **Proposal 6: The consistency/association of Set B beams and Set A beams across training and inference shall be guaranteed in RAN4 core requirement as the NW-side additional assistance information for UE-sided model testing for inference and monitoring.**  **Proposal 7: RAN4 requirement for model inference with UE-side AI/ML model needs to be specified, by considering the (potentially) new mechanisms, including (1) indication of the associated Set A from network to UE, (2) beam indication from network for UE reception (if new indication introduced), (3) predicted L1-RSRP(s) by new L1 signaling and (4) confidence/probability information.**  **Observation 5: For UE-sided model, RAN4 requirement on performance monitoring will be specified based on RAN1 conclusion on performance metric(s) and benchmark/reference for the performance comparison.**  **Proposal 8: For different types of performance monitoring for UE-sided model, the necessity of RAN4 requirement is provided as:**   1. **Type 1, Option 1 (NW-side performance monitoring): The necessity of RAN4 requirement on data collection for monitoring is not significant, because it is similar to data collection for other purposes.** 2. **Type 1, Option 2 (UE-assisted performance monitoring): RAN4 requirement on data collection for monitoring can be specified to test the accuracy of performance metrics calculated by UE.** 3. **Type 2 (UE-side performance monitoring): No RAN4 requirement is needed because no UE feedback will be performed.**   **Proposal 9: For LCM for UE-sided model, the necessity of RAN4 requirement is provided as:**   1. **LCM performed by the gNB: RAN4 requirement is needed.** 2. **LCM performed by the UE: No RAN4 requirement is needed.**   *Testability Issues for Testing AI-BM*  **Proposal 10: Testability study on testing AI-BM shall be based on existing FR2 OTA chamber system.**  **Proposal 11: The following testability questions shall be answered, if the Rel-15 NR testability SI compatible chamber is used for testing AI-BM:**  **Q1: With the limited number of AoAs NMAX\_AoAs = 2, is that possible to generate multiple beams from Set-A and Set-B?**  **Q2: The simulated multi-path fading propagation conditions for different DL TX beams can be correlated and logically correct (corresponding to a given beam book) to be learnt and inferred by AI/ML?**  **Q3: For BM-Case2, to test the temporal DL Tx beam prediction, the signals from DL TX beams can be predictable with continuity in time-domain?**  **Q4: For BM-Case2, the signal change for different DL TX beams can be correlated and logically correct (corresponding to a given beam book) to be learnt and inferred by AI/ML?** |
| [**R4-2402414**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402414.zip) | Ericsson | [Observation 1 For the type 1 performance monitoring and option 1, there may be a need to discuss requirements on the accuracy of reporting that is sent by the UE to the network. Potentially also measurement duration requirements may need to be considered.](#_Toc159255560)  [Observation 2 For the type 1 performance monitoring and option 2, it may be necessary to consider requirements on the accuracy of the reported performance metrics from the UE, in order that the network experiences comparable performance metrics from different types of UE operating a particular AI functionality.](#_Toc159255561)  [Observation 3 For the type 2 performance monitoring, requirements may be needed on the accuracy of reported performance metrics, in order that the network experiences comparable performance metrics from different types of UE operating a particular AI function. Alternatively, requirements relating to the event triggering may be needed if the UE triggers specific events based on the monitoring.](#_Toc159255562)  [Observation 4 For the type 2 performance monitoring, if the UE makes decisions on model selection/activation/deactivation/switching/fallback, then there may be a need for requirements that ensure that the UE makes proper decisions such that the performance is maintained, and also to ensure performance continuity.](#_Toc159255563)  [Observation 5 If the network indicates to the UE to do LCM operations, there may be a need for requirements such as activation/deactivation time, interruption time etc.](#_Toc159255564)  [Observation 6 For the network sided model, is a need to discuss L1 beam reporting enhancement measurements, in particular the accuracy of the beam reporting that is used for network inference.](#_Toc159255565)  [Observation 7 For the network sided model is a need to discuss the number of L1 beam measurements to be made by the UE and reported for NW sided inference, in particular whether the number of beams should be greater than four.](#_Toc159255566)  [Observation 8 For the UE sided model, a performance requirement may be related to the accuracy of AI inferred L1-RSRP. The accuracy of measured and inferred L1-RSRP may be differentiated in RAN4 requirements.](#_Toc159255567)  [Observation 9 For the UE sided model, when discussing requirements, consideration is needed on how the network sets the beams that are measured.](#_Toc159255568)  [Observation 10 If the UE reports confidence/probability information for AI model inference, RAN4 should discuss whether and how requirements and tests could be made for the accuracy of the confidence/probability information.](#_Toc159255569)  [Observation 11 If the UE reports best beams, the RAN4 requirements may be based around the accuracy of selecting the best beams.](#_Toc159255570)  [Observation 12 If data collection is within RAN, there may be a need for requirements on accuracy and timestamp accuracy of UE measurements.](#_Toc159255571)  [Observation 13 The amount of network assistance information for UE data collection should be minimized.](#_Toc159255572)  [Observation 14 If setting requirements on model inference performance for beams at different time instances, the range of applicable Doppler needs to be considered and operation across different Doppler speeds assessed.](#_Toc159255573)  [Observation 15 If defining requirements on NW sided inference, UE assistance information may need to be considered.](#_Toc159255574) |

## Open issues summary

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

1. KPIs – new metric? See Apple
2. Discuss measurement accuracy improvement? See SS paper Reuse legacy RSRP reporting for NW-side prediction
3. Test setup feasibility
4. Channel models in tests
5. Ground truth vs. UE measurements
6. Datasets to be used

### Sub-topic 2-1

*Metrics/KPIs for Beam management*

Different metrics/KPIs have been discussed and were captured in the TR:

For metrics for beam management requirements/tests, the following test metrics are identified and could be considered

- Option 1: RSRP accuracy

- Option 2: Beam prediction accuracy

- Top-1 (%) : the percentage of "the Top-1 strongest beam is Top-1 predicted beam"

- Top-K/1 (%) : the percentage of "the Top-1 strongest beam is one of the Top-K predicted beams"

- Top-1/K (%) : the percentage of "the Top-1 predicted beam is one of the Top-K strongest beams"

- Option 3: The successful rate for the correct prediction which is considered as maximum RSRP among top-K predicted beams is larger than the RSRP of the strongest beam – x dB,

- Related measurement accuracy can be considered to determine x

- Option 4: combinations of above options

RAN4 should continue to discuss what metrics are more appropriate and how they impact

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

Proposals

* + Option 1: Use Option 1
  + Option 2: Neither Option 1, 2, 3 is appropriate, a new metric is needed
  + Option 3: Use Option 2
  + Option 4: Combination of the above
  + Option 5: discuss new metrics
* Recommended WF
  + To be discussed

Companies suggesting to use different metrics(new) should come up with a concrete proposal

### Sub-topic 2-2

*Measurement accuracy*

Several companies brought up the need to improve measurement accuracy (e.g. RSRP measurement accuracy) to enable better AI/ML performance

**Issue 2-2: Measurement accuracy**

* Proposals
  + Option 1: RAN4 should consider tighter accuracy requirements for existing measurements (e.g. RSRP)
  + Option 2: RAN4 should study the impact of measurement accuracy on performance before discussing any possible tightening
  + Option 3: Accuracy requirements cannot be tightened
  + Option 4: Others
* Recommended WF
  + To be discussed

If option 2 is proposed, companies should also come up with more concrete proposals on how RAN4 should proceed

### Sub-topic 2-3

*Test setup feasibility*

Some companies brought up the issue of whether the FR2 test setup can accommodate the AI/ML tests, this is issue has to be discussed

**Issue 2-3: Test setup feasibility for FR12**

* Issues raised:
  + Option 1: **Testability study on testing AI-BM shall be based on existing FR2 OTA chamber system.**

**The following testability questions shall be answered, if the Rel-15 NR testability SI compatible chamber is used for testing AI-BM:**

* + **Q1: With the limited number of AoAs NMAX\_AoAs = 2, is that possible to generate multiple beams from Set-A and Set-B?**
  + **Q2: The simulated multi-path fading propagation conditions for different DL TX beams can be correlated and logically correct (corresponding to a given beam book) to be learnt and inferred by AI/ML?**
  + **Q3: For BM-Case2, to test the temporal DL Tx beam prediction, the signals from DL TX beams can be predictable with continuity in time-domain?**
  + **Q4: For BM-Case2, the signal change for different DL TX beams can be correlated and logically correct (corresponding to a given beam book) to be learnt and inferred by AI/ML?**
  + Option 2:
    - Based on the agreements in TR 38.843 on beam prediction quoted above, we propose to study the test feasibility for beam prediction from the following perspectives:
      * How to model the randomness and variation in time and spatial domain of L1-RSRP by the propagation conditions and the Tx beamforming gain on the AoDs considered in the propagation conditions
      * The verification procedures to confirm that the power level probability distributions on different AoAs aligns with the probability distributions derived from the configured propagation conditions and Tx beamforming gain
      * The number of beams and their coverage in spatial domain in Set B and Set A required to test beam prediction performance
    - We propose the following setup as the starting point for the feasibility studies
      * Consider CDL channel to model the randomness in the propagation conditions so that AoD, AoA and fading can be captured
      * Consider the antenna configuration from TR 38.901 clause 7.3 to design the beam peak directions, derive the corresponding beam patterns and determine the Tx beamforming gain on the concerned AoDs in the propagation conditions
      * Consider to define TE verification procedures to ensure that the power level probability distributions on different AoAs aligns with the probability distributions derived from the configured propagation conditions and Tx beamforming gain
      * Consider emulating UE rotation in the beam prediction test to model the AoA changes from UE perspective.
  + Option 3: Any other issues needed to be discussed
* Recommended WF
  + To be discussed

Test setup needs have to be discussed. Input from TE vendors would be highly appreciated

### Sub-topic 2-4

*Channel models in tests*

Some companies brought up the issue of what type of channel models to be used in tests

**Issue 2-4: Channel models**

* Proposals
  + Option 1: Fading channels are needed, RAN4 should study how to use them
  + Option 2: Only use AWGN channels
  + Option 3: Fading channels cannot be used because measurements cannot be checked
  + Option 4: Others
* Recommended WF
  + Option 1?

### Sub-topic 2-5

*Ground truth vs. UE measurements*

Some companies brought up the fact that it is difficult to check if UE measurements/reported value are correct relative to the signal input levels.

Ground truth definition also needs to be discussed.

**Issue 2-5: Ground truth vs. UE reported measurements**

* Proposals
  + Option 1: ground truth is the actual value that the UE should report (ideal value)
  + Option 2: Ground truth for a measured value cannot be established because of several reasons:
    - If fading is used, TE does not know the actual moment the UE takes a measurement sample
    - UE Rx gain is not known in the specific direction that the signal is coming from
    - Others
  + Option 2: Ground truth can be established known
  + Option 3: Others
* Recommended WF
  + To be discussed

### Sub-topic 2-6

*Data sets*

**Issue 2-6: Datasets for training/testing**

* Proposals
  + Option 1: Dataset based on channel models in TR 38.901
  + Option 2: RAN4 should discuss the feasibility of using dataset based on field data
  + Option 3: non-stationary channel modelling methods
  + Option 4:
* Recommended WF
  + To be discussed

# Topic #3: Testability and interoperability issues for positioning accuracy enhancement

This section contains the sub-topics regarding specific issues for positioning

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2400092**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400092.zip) | CATT | **Proposal 1: The positioning use case tests for performance, LCM and generalization should be defined separately.**  **Proposal 2: One general performance test procedure can be used for all AI/ML enabled positioning use case tests.**  **Proposal 3: The data for inference can be collected by DUT by measuring reference signals. The data type depends on the type required by the deployed AI/ML models/functionalities.**  **Proposal 4: For the data collected for inference, legacy measurement accuracy and reporting requirements can be reused. New accuracy requirements should be defined if new measurement quantities are defined.**  **Proposal 5: Regarding the data used for monitoring/verification purpose in direct AI/ML positioning tests, more RAN1 progress is required.**  **Proposal 6: Regarding the data used for monitoring/verification purpose in AI/ML assisted positioning tests, at least the legacy measurement quantities based on non-AI methods can be used as ground truth. Other options are not precluded. RAN4 needs to discuss how to verify the performance if new measurement quantities are defined.**  **Proposal 7: Quality indicator can be used to filter high-quality data, but more normative progress in RAN1 is required.** |
| [**R4-2400136**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400136.zip) | CAICT | **Proposal 1: Following issues are identified for specifying tests of direct positioning in RAN4:**   * **Construction of test environment** * **Acquisition of labels for model training** |
| [**R4-2400507**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400507.zip) | Apple | **Proposal 1: Positioning accuracy should be regarded as a test metric/KPI in RAN4 for the validation of AI/ML inference. RAN4 to further discuss the feasbility of testing this KPI. Positioning test data sets could be one option for testing this KPI.**  **Proposal 2: For Assisted AIML Positioning, the KPIs test metric (e.g., LOS/NLOS) needs to be considered for validating the positioning accuracy**  **Proposal 3 : In AI/ML-based positioning, it is essential to investigate the performance requirements for the input parameters of the positioning model/functionality (e.g., measurement accuracy of RSRP, ToA, RSRPP, RSTD) across all AI/ML positioning cases.**  **Proposal 4: In the context of AI/ML-based positioning, RAN4 shall explore the feasibility of testing KPIs such as RSRP, ToA, RSRPP, and RSTD. This exploration is crucial to ensure that the AI/ML model can be effectively and realistically trained.**  **Proposal 5: CIR/PDP should not be categorized as intermediate KPIs because they can only serve as inputs to AI/ML functionality/model and cannot function as outputs.** |
| [**R4-2401043**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401043.zip) | CMCC | ***Proposal 1: for direct AI/ML positioning, it is proposed to define requirements, and the metrics for positioning requirements/tests is proposed as positioning accuracy: Ground truth vs. Reported.***  ***Proposal 2: for AI/ML assisted positioning, it is proposed to discuss whether same metric(s) for requirements/tests can be used for both UE-side model and gNB-side model.***  ***Proposal 3: except ToA, RSTD and RSRP, and RSRPP, it is proposed to consider CIR/PDP as new measurement to define requirements.***  ***Proposal 4: when specify performance reqirements for AI/ML based positioning, it is proposed to discuss whether and how to consider the impact due to different assumption, e.g.model-input Size Reduction, non-ideal label(s), etc.*** |
| [**R4-2401611**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401611.zip) | vivo | ***Observation 1: Many issues need to be considered for RAN4 at least including: 1. How to obtain the ground truth value; 2. How to define requirements for metrics; 3. The generalization performance.***  ***Proposal 1: RAN4 to consider model inference output for direct AI/ML positioning only in case 1 in future discussion.***  ***Proposal 2: RAN4 to further discuss whether and how to define requirements and tests for model inference output for direct AI/ML positioning.***  ***Proposal 3: RAN4 to define requirements and tests for the intermediate features that already exist as the Rel-16/17 NR positioning measurements. Whether enhancements are needed for the requirements is FFS.***  ***Proposal 4: RAN4 to study how to define test and requirement for ToA.***  ***Proposal 5: Discussion about LOS/NLOS indicator could be deprioritized.***  ***Proposal 6: RAN4 to study whether and how to define requirements and tests for model input, including new potential measurements such as CIR/PDP.***  ***Proposal 7: RAN4 to define latency/interruption requirement for AI/ML based positioning based on RAN1/2 progress.***  ***Proposal 8: RAN4 to study how to test AI/ML based positioning under fading channel propagation conditions.*** |
| [**R4-2401686**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401686.zip) | Huawei,HiSilicon | ***Proposal 1:*** For UE-sided direct AI/ML positioning (Case 1), positioning accuracy is not testable, if the positioning result derived by UE is not reported to NW.  ***Observation 1:*** For Case 2a/2b and 3a/3b, RAN4 can start investigating the related testability after achieving sufficient input from other WGs.  ***Observation 2:*** For UE-assisted/NG-RAN node-assisted positioning with LMF-side positioning (Case 2a/3a), the relationship between measurement accuracy and positioning accuracy is unavailable, which has an impact on the test requirement definition. |
| [**R4-2401816**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401816.zip) | OPPO | **Proposal 1: 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.**  **Proposal 2: For case1, positioning accuracy should be utilized as the KPI(if feasible) to test the model/functionality output. If it is possible/how to obtain the positioning label needs to be considered and discussed in R19 with high priority.**  **Proposal 3: For case3a, 3b (cases without UE-side model), not necessary to test the Positioning model/functionality outputs.**  **Proposal 3: For case2a/2b, should be treated with 2nd priority in RAN4 R19.** |
| [**R4-2401819**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401819.zip) | ZTE Corporation | **Observation 1: RAN4 studies the one-side model with the UE-side not the gNB side.**  **Proposal 1: For case 1 and 2a, the model output can be the RSTD, identification of Los/NLos and other metric, the accuracy for these intermediate results could be considered for UE-side assisted positioning.**  **Proposal 2: For case 2b, 3a, 3b, RAN4 is not necessary to test the positioning model outputs in current stage since we study the one-side model with the UE-side not the gNB side.** |
| [**R4-2402305**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402305.zip) | Nokia, Nokia Shanghai Bell | In this paper we share our views on potential RAN4 impacts from issues related to AI/ML based positioning. Specifically, we cover following aspects:   * Selected sub use cases * Performance metrics * Measurement accuracy requirements * Testability aspects * LCM related aspects and performance monitoring   In the paper, the following Observations and Proposals were made:  **Observation 1:** Among 1st priority cases, only Case 1: UE-based positioning with UE-side model, direct AI/ML positioning may have impacts in RAN4.  **Proposal 1: RAN4 should initially prioritize Case 1 direct AIML positioning in the normative work.**  **Proposal 2: Positioning accuracy should be considered as the performance metric for case 1 - UE-based positioning with UE-side model, direct AI/ML positioning.**  **Proposal 3: In case any new measurement(s) are supported as AI/ML functionality input for Case 1 in RAN1, RAN4 may have impacts on measurement accuracy requirements.**  **Observation 2:** 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).  **Proposal 4: RAN4 should further study the feasibility of test mechanisms for positioning accuracy metric verification.**  **Observation 3:** If an LCM action is required and it is not taken in a timely manner, the performance degradation for AI/ML enabled Positioning use case may be degraded to undesirable level.  **Proposal 5: RAN4 core requirements should be considered to limit the latency of LCM actions towards the DUT for AI/ML enabled Positioning.**  **Proposal 6: RAN4 should follow the RAN1 discussions/ agreements on performance monitoring aspects for evaluation of any impacts in RAN4, particularly for Case 1 (UE-based direct positioning with UE-side model).** |
| [**R4-2402387**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402387.zip) | Samsung | **Proposal 1: RAN4 should be focused working on the three first prioritized cases by following the case prioritization in the WID although it could be changed by RAN or other WGs in the future based on the further study.**  **Observation 1: For data collection in training/finetuning, the data generation entity can be PRU/UE for case 1, TRP for case 3a, and TRP(input)/LMF(label) for case 3b;**  **Observation 2: For data collection in inference, the data generation entity is UE for case 1 and TRP for case 3a/3b;**  **Observation 3: For data collection in monitoring, the data generation entity is dependent on the monitoring metric.**  **Observation 4: RAN1 can support the collected data signalled from PRU (other UE) to UE deployed with AI model.**  **Proposal 2: RAN4 shall not define any requirement for data collection for UE-sided model training.**  **Observation 5: As a side information for the model output, the time stamp and quality information can be introduced.**  **Observation 6: TRP supports to signal the model output to LMF in case 3a.**  **Proposal 3: Before defining the requirement, the feasibility of model inference testing should be clarified for Case 1 at least. In addition, if Case 2a is confirmed to be introduced in RAN1, the feasibility of testing model inference for the case also needs the clarification.**  **Observation 7: RAN1 considers the model output based, model input based and other (measurement) based monitoring metrics.**  **Observation 8: RAN1 considers the monitoring entity as UE/TRP/LMF for UE side model, TRP, LMF for TRP side model, LMF for LMF side model.**  **Proposal 4: At least for UE-sided model, Case 1 and 2a, UE is prioritized as the entity to derive monitoring metric.** |
| [**R4-2402695**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402695.zip) | Ericsson | In this contribution RAN4 issues related to AI/ML based positioning are discussed. Since this is the first contribution to rel. 19 AI/ML based positioning normative work, this contribution emphasizes on Ericsson’s view on the use cases and the associated requirements that fall within the RAN4 scope. The discussion presented in this paper can be summarized into the observations and proposal listed below.  **Observation 1**: Use case 1 is equivalent to UE-based positioning in legacy NR positioning.  **Observation 2**: Requirements are not defined for measurements for UE-based positioning in legacy NR positioning specification. In this regard, discussion on core and performance requirement for use case 1 is not relevant for RAN4 discussion.  **Proposal 1**: RAN4 to not define requirements for use case 1.  **Observation 3**: Use case 2a is similar to legacy UE-assisted NR positioning, in the sense that the measurements are generated by the AI/ML model in UE and the measurements are reported to LMF where UE position is estimated.  **Observation 4**: RAN4 has defined requirements for all the measurements associated to UE-assisted positioning.  **Observation 5**: Use case 2a is relevant for core and performance requirements discussion in RAN4.  **Proposal 2**: Define core and performance requirements for use case 2a. Details and the scope of the core and performance requirements for use case 2a depends on the outcome of RAN1/RAN2 discussions.  **Observation 6**: For case 2b model input are measured by UE and transferred over LPP to LMF.  **Observation 7**: Accuracy of model inferencing for use case 2b relies on the accuracy of measurement performed by the UE.  **Proposal 3**: RAN4 to define core and performance requirements for use case 2b.  **Proposal 4**: RAN4 to define performance requirement for use case 3a. Details of the applicable performance requirement are identified based on the progress made by RAN1 on the use case.  **Observation 8**: For case 3b model input are measured by gNB and reported to LMF over NRPPa.  **Proposal 5**: RAN4 to define performance requirement, limited to measurement performed and reported by gNB, for use case 3b.  **Observation 9**: Framework for AI/ML model training data generation is not clear enough to identify the RAN4 impact. |

## Open issues summary

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

1. Requirements for case 1
2. Requirements for case 3a/3b
3. Accuracy requirements for existing measurements/reports
4. KPIs for case 1
5. KPIs for case 3a/3b

### Sub-topic 3-1

*Requirements for case 1*

There are some proposals on whether to define requirements for case 1 (UE-based positioning with UE-side model, direct AI/ML positioning)

**Issue 3-1: Requirements for case 1**

* Proposals
  + Option 1: RAN4 should not define requirements for case 1
  + Option 2: RAN4 should continue to discuss how to define requirements for case 1 (including feasibility of defining such requirements)
  + Option 3: Others
* Recommended WF

To be discussed

Note: currently there are no requirements for UE based positioning

### Sub-topic 3-2

*Requirements for case 3a/3b*

Requirements for case 3a/3b would be defined on some network nodes/entities. Such accuracy requirements have not been defined before

**Issue 3-2: Requirements for case 3a/3b**

* Proposals
  + Option 1: RAN4 should not defined positioning accuracy requirements for case 3a/3b (accuracy on inference results of AI/ML positioning model)
  + Option 2: RAN4 should continue to discuss how to define requirements for case 3a/3b (including feasibility of defining such requirements)
  + Option 3: RAN4 continues to discuss requirements on measurements reported by UE/gNB to the entity running the AI/ML positioning model
  + Option 4: others
* Recommended WF
  + To be discussed, options are not exclusive

### Sub-topic 3-3

*Measurements and reported metrics/values*

**Issue 3-3: Handling of requirements for measurements and reported metrics/values**

* Proposals
  + Option 1: Reuse the already defined requirements for existing measurements/reported metrics, RAN4 should only discuss new requirements if new metrics to be measured/reported are introduced by other groups
  + Option 2: RAN4 should look into enhancing/tightening existing requirements
  + Option 3: others
* Recommended WF
  + Option 1

### Sub-topic 3-4

*KPIs for case 1*

Several KPIs are proposed for each use case, it should be discussed which are more appropriate for each use case if requirements are to be studied/defined

Identified KPIs in the TR:

For metrics for positioning requirements/tests, the candidate options include

* Option 1: positioning accuracy: Ground truth vs. reported
  + only option available for direct positioning
* Option 2: CIR/PDP, channel estimation accuracy
* Option 3: ToA, RSTD and RSRP, and RSRPP
* Option 4: others (e.g., intermediate KPIs, LoS/NLoS)/combinations of the above

**Issue 3-4: KPIs for case 1**

* Proposals
  + Option 1: Option 1
  + Option 2: Option 2
  + Option 3: Option 3
  + Option 4: others
  + Option 5: combination of options
* Recommended WF
  + To be discussed

If option 5 is proposed, a concrete proposal should be presented

### Sub-topic 3-5

*KPIs for case 3a/3b*

Several KPIs are proposed for each use case, it should be discussed which are more appropriate for each use case if requirements are to be studied/defined

Identified KPIs in the TR:

For metrics for positioning requirements/tests, the candidate options include

* Option 1: positioning accuracy: Ground truth vs. reported
  + only option available for direct positioning
* Option 2: CIR/PDP, channel estimation accuracy
* Option 3: ToA, RSTD and RSRP, and RSRPP
* Option 4: others (e.g., intermediate KPIs, LoS/NLoS)/combinations of the above

**Issue 3-5: KPIs for case 3a/3b**

* Proposals
  + Option 1: Option 1
  + Option 2: Option 2
  + Option 3: Option 3
  + Option 4: others
  + Option 5: combination of options
* Recommended WF
  + To be discussed

If option 5 is proposed, a concrete proposal should be presented

# Topic #4: Testability and interoperability issues for CSI compression and CSI prediction

This section contains the sub-topics regarding CSI compression and prediction

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2400093**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400093.zip) | CATT | **Proposal 1: The test cases for performance, LCM and generalization for both CSI prediction and compression use cases can be defined separately.**  **Proposal 2: The data collected for performance verification in the CSI prediction tests can consider the following data:**  **- Performance metric(s), e.g., SGCS/NMSE.**  **- Performance monitoring output (Need more RAN1 discussions).**  **- Predicted CSI and/or the corresponding ground-truth (Need more RAN1 discussions).**  **Proposal 3: Our views on the comparison of the four options of test decoders are provided in the table. And we are okay to drop Option 1 and Option 2 and focus on discussing Option3/4.** |
| [**R4-2400135**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400135.zip) | CAICT | **Proposal 1: Our views on two-sided testing options comparison are listed in the table with additional notes. (see Annex)**  **Proposal 2: If CSI prediction accuracy is taken as metric to test inference performance, the measured CSI on DUT-side could be considered as reference to calculate the accuracy.**  **Proposal 3: Beside throughput, other candidate options (e.g., applicable condition, dataset characteristic and etc.) could be considered for discussion on monitoring metrics.** |
| [**R4-2400508**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400508.zip) | Apple | **Observation 1: To bound the variations in the TE implementations for the encoder/decoder for option 4, some parameters of the module structure should be specified by RAN4.**  **Observation 2: There are multiple issues with the adaptation of a standardized training dataset. The issues range from providing multiple training datasets, to dataset obsolescence and generalizability concerns in real deployment.**  **Observation 3: There are performance issues if the training of test decoder 4 is done with type 3 training. The performance loss depends on the discrepancy between the backbones of encoder and decoder, and if the training dataset consists of a diverse dataset from multiple trained encoders. There is currently no definitive answer regarding how or what elements must be captured to attain reliable performance across all deployment scenarios.**  **Proposal 1: The following parameters may be specified for test decoder for option 4 by RAN4:**   * **Model type (backbone: CNN, Transformer, MLP, etc)** * **Model structure (depth, width, etc)** * **Activation functions** * **Maximum FLOPs allowed for the test decoder** * **Maximum number/size of parameters**   **Proposal 2: RAN4 should investigate the adoption of a more adaptive strategy towards a standardized dataset and explore a collaborative approach involving multiple entities to gather training data for the test decoder targeting option 4.** |
| [**R4-2400562**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2400562.zip) | Qualcomm, Inc. | **Proposal 1: To design the fully specified decoder, RAN4 follows the procedures below under the assumption that basic setup including propagation conditions and UE/CSI configurations were agreed:**   1. **RAN4 agrees the encoder input data generation procedure(s) and encoder/decoder pair backbone structure(s)** 2. **RAN4 agrees the cost function for training purposes** 3. **The parameters of the encoder/decoder pair are trained based on the generated input data, backbone structure and the cost function**   **Note: whether to specify further implementation details, e.g., fixed point design for the neural network implementation, can be decided by assessing the performance impact after agreements are reached in the above three steps.**  **RAN4 can start with collecting proposals of encoder input data generation procedure(s) and encoder/decoder pair backbone structure(s), and we provide an example below:**   * **Encoder input data generation procedure: start with nominal channel estimation procedures followed by singular value decomposition to derive the precoders as the encoder input, if this procedure provides a good coverage of encoder input space of different UE implementations.** * **Encoder/decoder pair backbone structure: a simple fully connected neural network (further details can be included in the proposals when RAN4 is deciding the backbone structure to minimize the variation in step (2))**   **The propagation condition and CSI configurations should be discussed in WI phase. For the test decoder design discussion purpose, we can start with the scenario defined in 38.101-4 “*6.3.2.2.6 Single PMI with 16Tx Enhanced Type II Codebook*”.**  **Proposal 2: To design the partially specified decoder, RAN4 starts with the procedures in proposal 1 to derive the reference encoder/decoder pair for dataset generation or decoder verification purposes. Note that option 4a-2 can accommodate multiple reference encoder/decoder pairs, while option 4a-1 and 4b require RAN4 to agree to one pair of encoder/decoder. Based on the encoder input generation procedure and reference encoder/decoder pair agreements, RAN4 has the following options to (partially) specify the test decoder:**   * **Option 4a: Capturing a {encoder input, encoder output, decoder output} dataset in RAN4 specification, and the test decoder implementations are verified against this {encoder output, decoder output} dataset. Two sub-options for dataset generation are listed below:**   + **Option 4a-1: The dataset is generated by one agreed reference encoder/decoder pair (for dataset generation purpose)**   + **Option 4a-2: The dataset is generated by the encoder/decoder pairs designed by the contributing companies based on the agreed common assumptions** * **Option 4b: Capturing the encoder in the agreed reference encoder/decoder pair (for test decoder verification purpose) in RAN4 specification, and the test decoder implementations are verified against this reference encoder.**   **Proposal 3: RAN4 studies the potential solutions to the following open issues for options in proposal 2:**   * **Option 4a-1: whether it is feasible to design a standardized dataset sufficiently representing the propagation channel condition and possible UE procedures (channel estimation, whitening and desired precoder derivation).** * **Option 4a-2: whether it is feasible to design a test decoder that can properly recover the encoder input from the encoder output, given that the dataset is from multiple encoder and decoder pairs**. * **Option 4b: whether it is feasible and how to define a proper test data generation procedure to guarantee test result consistency by the verification procedure with the reference encoder.**   **Proposal 4: Supported training type discussion is deferred to WI stage.**  **Proposal 5: “*reflection on the real deployment*” can be discussed in following perspectives:**   * **Knowledge of the test decoder** * **Whether the decoder vendor can implement the test decoder** * **Do not discuss the training type perspective if the clarification of training types is deferred to WI stage**   **Proposal 6: The reference encoder (w.r.t. the test decoder) and the reference decoder (w.r.t. the test encoder) specification discussion for simulation alignment purpose follows the principle:**  **During the simulation assumption discussion in WI stage, RAN4 can choose one of the following options of reference encoder/decoder pair based on the alignment across collected preliminary simulation results**   * + - * **Option 1: Apply agreed common assumptions**       * **Option 2: Introduce additional simulation assumptions in addition to agreed common assumptions**       * **Option 3: (fully) specify the reference encoder/decoder pair**   **In addition, based on the options in proposal 2, we may need the reference encoder/decoder pair for dataset generation or decoder verification purposes.**  ***See Table in the Annex***  **Proposal 7: Option 1 and 2 are not feasible due to confidentiality issues of sharing proprietary information/models between different companies.**  **Proposal 8: Option 3 is feasible since there is a proper decoder derivation procedure as described in proposal 1 and relatively low TE requirements.**  **Proposal 9: Option 4 is feasible due to lowest TE requirement. The two additional properties, test repeatability and implement ability of other vendors, can be achieved by the options presented in proposal 2.** |
| [**R4-2401045**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401045.zip) | CMCC | ***Proposal 1: for CSI compression and CSI prediction, it is proposed to use intermediate KPI as requirements/tests metrics for LCM.***  ***Proposal 2: In oder to make decision/selection of test decoder/encoder for 2-sided model, even though many aspects are considered in the comparison of options in Rel-18, it is proposed to prioritize some of the aspects, e.g, reflection on the real deployment and specification effort can be prioritized.*** |
| [**R4-2401172**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401172.zip) | xiaomi | See Table with comparison in the Annex |
| [**R4-2401612**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401612.zip) | vivo | **Observation 1: Model structure (back-bone, parameters, e.g., number of layers, etc) also has significant performance impact even if complexity of model (in terms of FLOPS) are similar.**  **Observation 2: 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 for both 1-sided model and 2-sided model.**  **Proposal 4: In Option 3, model structure needs to be specified first, and then model parameters can be further specified.**  **Proposal 5: In Option 4, at least model structure and training dataset need to be specified. To make alignment of model structure, model backbone and model complexity (e.g., model (parameter) size and FLOPs) could be aligned first.**  **Proposal 6: The suggested model structures in Figure 2.4-1 to 2.4.4 for test decoder/encoder could be used as a starting point**  **Proposal 7: “Supported training collaboration type between DUT and decoder provider” can be removed from the table of the comparison of the four options of test decoder, since this aspect is just for training before test and seems to have no obvious impact on the test.**  **Proposal 8: Take into consideration the summary of 4 options for testing of 2-sided model in Table 2.4-1.**  **Proposal 9:** **Throughput/relative throughput based solution would be the baseline for CSI prediction. The** γ **value of predicted PMI vs random PMI can be used in the test, which would be different than the** γ **value of previous non-AI test.**  **Observation 3: Post deployment performance may be verified by model monitoring.**  **Proposal 10: RAN4 to further study whether and how to test the post-deployment AI/ML feature for ensuring the performance in Rel-19 stage.** |
| [**R4-2401687**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401687.zip) | Huawei,HiSilicon | ***Proposal 1***: Deprioritize Options 1 and 2 for determining the test decoder of two-sided model.  ***Proposal 2***: According to whether using a mixed training dataset to determine the reference decoder, Option 3 can be further divided into two sub options as follows.   * Option 3a: The test decoder is determined for each test case, using a specific dataset collecting from the configuration/scenario of the considered test case. * Option 3b: The test decoder is determined for more than one test cases, using a mixed dataset collecting from different configurations/scenarios of the considered test cases.   ***Proposal 3***: Take Option 3a as baseline, where a specific rather than a mixed dataset is used for defining the test decoder in each test case.  ***Proposal 4***: For achieving a converged test decoder in Option 3, at least the structure of both the reference encoder and test decoder, hyperparameters of model training, as well as a determined sample-by-sample dataset are expected to be aligned among companies.  ***Observation 1***: Even with all hyperparameters aligned and model training converged, the model parameters provided by companies can still be different. How to align model parameters of the test decoder among companies is an open issue.  ***Proposal 5***: According to whether the model structure is specified, Option 4 can be further divided into two sub options as follows.   * Option 4a: Model structure is not specified in RAN4. Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder. * Option 4b: Model structure is specified in RAN4. Training dataset is not specified for verifying the encoder at DUT. The test decoder developed by TE vendor needs verification.   + FFS: How to determine the test metric for test decoder developed by each TE vendor.   ***Observation 2***: The boundary between Option 3 and Option 4b is whether the model parameters are specified in RAN4.  ***Proposal 6***: In Option 4, the performance of test decoder should be verified before testing DUT in each test.  ***Proposal 7***: Compression ratio and quantization level needs to be specified in Options 3 and 4.  ***Observation 3***: Though there is no need to align the model parameters of test decoder, model parameters of the reference encoder for verifying the test decoder is still needed. How to align is still an open issue.  ***Proposal 8***: The comparison of the four options of test decoder is updated as below. – see annex for table  ***Proposal 9***: Deprioritize SCGS/NMSE for defining baseline requirements in AIML-enabled CSI compression.  ***Proposal 10***: Deprioritize CSI prediction accuracy for defining baseline requirements in AIML-enabled CSI prediction.  ***Observation 4***: How to ensure that the testing dataset aligns well with training dataset is still an open issue. |
| [**R4-2401817**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2401817.zip) | OPPO | **Observation 1: Pros and cons for different options on test decoder are shown in table 1. – see annex**  **Proposal 1: Regarding the supported training collaboration types of different test options, the impact of different training collaboration types on the provider of encoder and decoder in practical use should be taken into account, and evaluate whether different testing options can be utilized accordingly to better align with actual use cases under such impact.**  **Proposal 2: Option3/4 could be baseline(if the test model could be specified or partial specified). Option 1 could be considered as an optional solution(for cases test model provided by DUT could be accepted by UE and TE vendor ).**  **Proposal 3: For both CSI compression and CSI prediction, model/functionality input (CSI-RS measurement accuracy) and output (associated throughput) related tests should be supported.**  **Proposal 4: 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 5: Reuse the legacy PMI requirement (compare to random precoding) as a baseline test, and let other options/proposals with higher performance requirements be further studied**  **Proposal 6: In R19, static test scenarios and configurations should be considered first. After having feasible testing cases for static configurations, then further consider whether to introduce non-static testing scenarios and configurations.**  **Proposal 7: The CDL/[UMa] channel can be prioritized as the baseline for static scenario/configuration testing** |
| [**R4-2402306**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402306.zip) | Nokia, Nokia Shanghai Bell | In this paper we share our views on potential RAN4 impacts from issues related to AI/ML CSI compression and CSI prediction. Specifically, we cover following aspects:   * Interoperability and training * Test decoder design options * Test metrics * On encoder/decoder pair for requirements and tests * LCM Core functionality management requirements * Generalization for CSI feedback   In the paper, the following Observations and Proposals were made:  **Observation 1:** CSI compression is the only two-sided AI/ML use case. addressing the Rel-18 challenge with performance vs complexity trade-off, temporal dimension will be studied in Rel-19. Inter-vendor collaboration and interoperability will also stay as the main challenges.  **Proposal 1: RAN4 should wait for further progress on RAN1-side in respect to temporal dimension introduction in CSI-compression to proceed with the use-case specific requirements and tests.**  **Observation 2:** CSI prediction use-case is provisioned by one-sided models and should follow the general RAN4 requirements and testing framework. Use-case specific aspects and requirements shall be still considered in Rel-19.  **Observation 3:** No inter-operability issues for the CSI-prediction are expected because a similar mechanism was already introduced in terms of Rel-18 NR\_MIMO\_evo\_DL\_UL WI.  **Proposal 2: RAN4 shall align model training collaboration to be agreed in RAN1 with the formulation of requirements and test decoder design.**  **Proposal 3: RAN4 to complete the missing cells in Comparison of the four options of test decoder table as proposed below.**  **Proposal 4: RAN4 should clarify the actual input that is expected from the row “Applicability to different scenarios/conditions/ configurations” so that it helps us to align the responses properly.**  **Observation 4:** Introduction of AI/ML-enabled CSI use cases will cause new UE performance requirements in TS 38.101-4. A new target value of γ (gamma) for AI/ML-enabled CSI use cases can be envisaged.  **Observation 5:** A new relative throughput performance metric may be introduced for AI/ML-enabled CSI use cases.  **Proposal 5: RAN4 should further study if a modified relative throughput metric based on actual/non-random non-AIML CSI feedback as a reference would be more suitable for performance requirements on AI/ML compressed CSI.**  **Proposal 6: Intermediate KPIs (such as SGCS, NMSE, Option 2 in Clause 7.3.3 from TR 38.843) may be still considered for performance monitoring of AI/ML-based CSI compression requirements, if they are introduced by RAN1 design.**  **Observation 6:** For CSI prediction performance monitoring, RAN1 is already discussing NW-side mechanism based on UE-reported ground truth. The same approach can be applied to test the CSI prediction accuracy as a performance KPI in RAN4.  **Observation 7:** CSI Prediction Accuracy as a KPI can be testable using the existing test interfaces with minimum change in the TE.  **Proposal 7: RAN4 should further discuss different options available to test the CSI prediction accuracy to ascertain the feasibility of defining CSI prediction accuracy as a Metric in the WI phase.**  **Proposal 8: CSI prediction accuracy metric (Option 3 in Clause 7.3.3 from TR 38.843) may be still considered for performance monitoring of AI/ML-based CSI prediction use case requirements, if they are introduced by RAN1 design.**  **Proposal 9: RAN4 should follow the same principles for test encoder/decoder specification and derivation of requirements, e.g., if Option 3/4 is agreed for the test decoder, then the same decoder/rules should be used for the requirements.**  Table 2: Test decoder specification parameters for Option 3   |  |  |  | | --- | --- | --- | | **Category** | **Parameter** | **Description/Examples** | | Model architecture parameters | Model type | Transformer, CNN, RNN, etc. | | Model depth | Number of layers | | Layer type | Fully connected, convolutional, etc. | | Layer size | Neuron count and configuration | | Quantization method | Scalar, vector (with codebook) | | Model Training related parameters | Training procedure | Initialization method, training duration, training completion criteria | | Loss function | SGCS, NMSE, etc. | | Hyperparameters | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc. | | Cross-validation details | Dataset splits for training/testing/validation | | Generalization (may be applicable to all four options) | Performance requirements on test dataset(s) | Mean SGCS, etc. | | Scalability (may be applicable to all four options) | Supported antenna port configurations | (2,8,2), (2,4,2), etc. | | Supported feedback payloads | Low, medium, high overhead (with specified number of bits) |   **Proposal 10: RAN4 should consider model architecture parameters and model training related parameters, as shown in Table 2 above, to fully specify test decoder Option-3.**  **Proposal 11: Regarding Option 4 of test decoder, RAN4 needs to discuss further which parameters can be standardized and which parameters can be left for implementation.**  **Observation 8:** If required LCM action is not taken in a timely manner, the performance of AI/ML-based CSI feedback may be degraded to undesirable levels.  **Proposal 12: Use-case specific core requirements should be considered to limit latency of LCM actions (e.g. activation, deactivation, fallback, switching etc.) typical for the CSI feedback enhancement use case.**  **Observation 9:** Scalability should be possible to treated with usual RAN4 approaches. However, for generalization aspects some additional information about the functionality need to be shared, which is currently unknown and would depend on RAN1 design. |
| [**R4-2402390**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402390.zip) | Samsung | *General RAN4 Scope for AI-CSI*  **Observation 1: It is out of RAN4’s scope to identify the performance gain of any particular use case which could be used as justification to confirm/disconfirm this use case to be included in the normative work in Rel-19.**  **Proposal 1: The RAN4 work scope related to CSI feedback enhancement shall only be related to the testability and interoperability objectives on two-sided model, at least before any CSI use case is confirmed at the checkpoints in RAN#105 (Sept ’24).**  *Interoperability and Testability issues for CSI compression use case:*  **Observation 2: As long as training data set (including enough amount of data for raw CSI and compressed bit strings) available in 3GPP standard for TE decoder training, it is possible for both TE and DUT vendors to develop test decoder and interoperable encoder.**  **Observation 3: If the training data set (including enough amount of data for raw CSI and compressed bit strings) is available in 3GPP standard, Option 4 can be regarded as the standardized training data set, compared with Option 3 with standardized test decoder itself.**  **Proposal 1: For Option 4,**  **- It is assumed that TE vendor will not share decoder to other vendors (DUT and/or infra vendors);**  **- Parameters that need to be specified for defining test decoder shall include:**  **🡺 Training data set for TE decoder training, including enough amount of data for raw CSI and compressed bit string.**  **Proposal 2: The following clarification of options are provided for option 1-4 test decoder for 2-sided model (except the green highlighted part for existing agreement).**  **See Table in the Annex**  **Observation 4: Among 4 options to derive test decoder, the feasibility of offline training for DUT vendor to obtain encoder can be confirmed for Option 1, 3 and 4 (where Option 4 needs further clarification).**  **Observation 5: 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 test model for two-sided model test implementation.**  **Observation 6: Among 4 options to derive test decoder:**  **- 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 offline collaboration is available between UE and gNB vendors) can be regarded to match with Type-2 or Type-3 training collaboration.**  **- Option 3 don’t need any training collaboration since test decoder is fully specified.**  **- Option 4 don’t need any training collaboration, but the training by DUT vendors is similar to Type-3 training collaboration.**  **Proposal 3: To further study the testability and interoperability of four options to derive test decoder:**  **- Option 1 shall be precluded because it can’t reflect any real deployment;**  **- Option 2 shall be precluded because it involves extensive effort required for multiple test decoders by different gNB vendors, and the required additional burden for decoder verification;**  **- Option 3 and 4 (with training data set specified for TE decoder training) can be adopted.**  *Discussion on CSI prediction use case:*  **Observation 7: For CSI prediction use case, there is no use case-specific issues on testability and interoperability identified.**  **Proposal 4: RAN4 start the discussion on CSI prediction-specific issue only after RAN-P confirmed this use case for normative work.** |
| [**R4-2402413**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110/Docs/R4-2402413.zip) | Ericsson | [Observation 1 Achieving interoperability is critical for the Uu interface.](#_Toc159255310)  Based on the discussion in the previous sections we propose the following:  [Proposal 1 Options 3 and 4 are the most obvious ways to achieve interoperability. The details of how they should work should be studied further based on the questions below.](#_Toc159255313)  For option 3, the following aspects need to be discussed in order to be able to create a standard:   * Should the standardized model be a RAN4 model (used for testing, but will implicitly determine the models needed at gNB and/or UE) ? Or should RAN1 standardize one or both sides models ?   + This question likely needs to be decided in collaboration with RAN1 * What architecture should be assumed for the test model, or even whether more than one model with more than one architecture should be standardized. * What size the model should have. * How the model should be trained; how can training data be made available ? * Is it possible to train a single model between interested companies (possibly by passing the model between interested vendors so that each could further train the model using their own proprietary data set) * If there is a need to decide between models, how to decide ? The model with the best performance ? * How to test the model ? Is both a standardized test decoder and test encoder needed in the end ? * By what means to physically capture the model in the specification ?   For option 4, the following needs to be discussed:   * What as a minimum needs to be standardized to, in effect capture the latent space of the encoder ? * Is there a way to avoid capturing the training dataset ? * Is there a need to capture a means to validate that a model trained using the specification achieves the expected performance ? Is a reference encoder (or reference decoder) needed ? * How to relate the captured data to the implied minimum complexity of a model derived from the specification ? |

## Open issues summary

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

1. CSI prediction accuracy
2. Test Encoder/decoder options for 2-sided model (drop option 1/2)
3. Option 3 for 2-sided model
4. Option 4 for 2 sided model
5. Comparison table

### Sub-topic 4-1

*CSI prediction accuracy metrics*

**Issue 4-1: CSI Prediction Accuracy metrics**

* Proposals
  + Option 1: Prediction accuracy can be used as KPI/metric
  + Option 2: Prediction accuracy cannot be used because the “correct” value is not available
  + Option 3: Throughput should be the default metric, others should be discussed only if throughput is not feasible
  + Option 4: Others
* Recommended WF

Option 3

### Sub-topic 4-2

*Test Encoder/decoder options for 2-sided model*

Some companies proposed to drop Options 1/2 from the options considered for 2-sided model due to confidentiality issues

**Issue 4-2: Testing options for 2-sided model**

* Proposals
  + Option 1: Do not consider options 1/2 further, only consider options 3/4
  + Option 2: Keep all options
  + Option 3: others
* Recommended WF
  + Option 1

### Sub-topic 4-3

*Option 3 for 2-sided model*

Several companies brought proposal on how to further study/check the feasibility of option 3

**Issue 4-3: Option 3 for 2-sided model**

* Proposals for parameters that should be agreed upon:
  + Option 1: Propagation conditions, UE/CSI configurations
  + Option 2:
    - encoder input data generation procedure(s) and encoder/decoder pair backbone structure(s)
    - cost function for training purposes
    - The parameters of the encoder/decoder pair are trained based on the generated input data, backbone structure and the cost function
  + Option 3:
    - Quantization, fixed point design
  + Option 4:
    - Activation functions
  + Option 5:
    - Maximum FLOPS
  + Option 6:
    - Number of layers, layer type, layer size
  + Others
* Recommended WF
  + To be discussed

Likely multiple options need to be chosen, RAN4 should agree on a minimum set such that companies can continue the study

### Sub-topic 4-4

*Option 4 for 2-sided model*

Several companies brought proposal on how to further study/check the feasibility of option 4. This discussion is also related to sub-topic 4-3 as the feasibility study has many similarities

**Issue 4-4: Option 4 for 2-sided model**

* Proposals
  + Option 1:
* **Capture a {encoder input, encoder output, decoder output} dataset in RAN4 specification, and the test decoder implementations are verified against this {encoder output, decoder output} dataset. Two sub-options for dataset generation are listed below:**
  + **Option 4a-1: The dataset is generated by one agreed reference encoder/decoder pair (for dataset generation purpose)**
  + **Option 4a-2: The dataset is generated by the encoder/decoder pairs designed by the contributing companies based on the agreed common assumptions**
  + Option 2:
* **Capturing the encoder in the agreed reference encoder/decoder pair (for test decoder verification purpose) in RAN4 specification, and the test decoder implementations are verified against this reference encoder.**
  + Option 3: discuss which parameters/assumptions should be different/not needed compared to Option 3 in Issue 4-3
  + Option 4: Model structure is not specified in RAN4. Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder.
  + Option 5:
* Model structure is specified in RAN4. Training dataset is not specified for verifying the encoder at DUT. The test decoder developed by TE vendor needs verification.
  + FFS: How to determine the test metric for test decoder developed by each TE vendor.
  + Option 6: Others
* Recommended WF
  + To be discussed

Likely multiple options need to be chosen, RAN4 should agree on a minimum set such that companies can continue the study

### Sub-topic 4-5

*Comparison table*

The comparison table has been discussed for several meetings, companies brought proposals on how to fill in the parts of the table which were not finalized during the SI phase. The table with the agreements so far highlighted in green is captured below. The companies’ inputs are included in the Annex.

**Issue 4-5: Comparison table**

* Proposals
  + Option 1:
* **Table 7.4.2.3-1 Comparison of the four options of test decoder**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendor (no need to be specified) | Up to decoder implementer (infra vendor)  FFS whether coordination with encoder vendor is required | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on the RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) |  |  |  |  |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | - Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  - Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long as the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used in actual field deployments ) |  |  |  |  |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower than Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/interoperability between TE and DUT needs further study. |
| Specification Effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs study and decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure(after specs are published) |  |  | No | No |
| Applicability to different scenarios/conditions/ configurations |  |  |  |  |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than option 1. | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than option 3/4  FFS compared to option 2 | Higher than Option 3/4  FFS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem |  |  |  |  |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented |  |  |  |  |
| Relationship with reference decoder/encoder(used by RAN4 to define the performance requirements) for defining requirement |  |  |  |  |
| Whether model transfer/delivery is needed during the test procedure |  |  |  |  |

* Recommended WF
  + Continue to discuss based on companies inputs.

# Annex: Companies’ Inputs on the Comparison of the four options of test decoder

R4-2400093 – CATT:

Table 7.3.2.3-1: Comparison of the four options of test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendors (no need to be specified) | Up to decoder implementer (infra vendor) | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | CATT:  Type 1/3;  Note1: Type 3 requires shared dataset and/or other information between UE and gNB vendors. | CATT:  Type 1/2/3;  Note1: Type 2 needs collaboration between UE and NW vendors. (Not preferred.)  Note2: Type 3 requires shared dataset and/or other information between UE and gNB vendors. | CATT:  Type 1/3;  Note1: Type 3 requires shared dataset and/or other information between UE and gNB vendors. | CATT:  Type 1/3;  Note1: Type 3 requires shared dataset and/or other information between UE and gNB vendors. |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long a the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS  CATT: A question here. How to verify the decoder performance on TE? An encoder required? Then we may fall into a loop.  If no feasible method found, our answer to this issue is ‘Not feasible’. | FFS  CATT: Same question as the one in the left. | FFS  CATT: Feasible. RAN4 can define a pair of encoder and decoder in spec. | FFS  CATT: Feasible. RAN4 can define a pair of encoder and decoder in spec. |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used | CATT:  Difficult to reflect the real deployment. | CATT:  Can reflect the real deployment. | CATT:  Depends on the encoder/decoder design. Difficult to make a conclusion at current stage. | CATT:  Depends on the encoder/decoder design. Difficult to make a conclusion at current stage. |
| TE requirements to deploy the decoder (e.g., training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower thank Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/ interoperability between TE and DUT needs further study |
| Specification effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs to study and may decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure (after specs are published) | CATT:  Probably involve confidentiality/IP issues between DUT and TE. | CATT:  Probably involve confidentiality/IP issues among DUT, decoder provider and TE. | No | No |
| Applicability to different scenarios/conditions/ configurations | CATT:  Applicable. | CATT:  Applicable. | CATT:  Depends on the encoder/decoder design and training dataset.  Applicable if training dataset is updated. | CATT:  Depends on the encoder/decoder design and training dataset.  Applicable if training dataset is updated. |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than Option 1 | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than Option 3/4  FSS compared to Option 2 | Higher than Option 3/4  FSS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | CATT:  High than Option 3/4.  Need for interaction between DUT and UE/gNB side in deployment. | CATT:  Low.  Both UE and gNB sides are involved during the tests. No extra work required in deployment. | CATT:  Medium.  Though UE/gNB is not involved during the tests, the encoder/decoder is public. | CATT:  Medium.  Though UE/gNB is not involved during the tests, the encoder/decoder is public. |
| Friendly to STOA (state of the art) model test / Forward compatibility when new AI models are invented | CATT:  Friendly (Only DUT involved). | CATT:  Not friendly (the other side is also involved). | CATT:  Friendly. | CATT:  Friendly. |
| Relationship with reference decoder/encoder (used by RAN4 to define the performance requirements) for defining the requirement | CATT:  Probably no relation at all. | CATT:  Probably no relation at all. | CATT:  Could be the same decoder/encoder. | CATT:  Could be the same decoder/encoder. |
| Whether model transfer/delivery is needed during the test procedure | CATT:  Need since TE need to load the decoder provided by DUT. | CATT:  Need since TE need to load the decoder provided by the decoder provider. | CATT:  Not needed. | CATT:  Not needed. |

R4-2400135:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of test decoder training data | Up to DUT vendor (no need to be specified) | Up to decoder implementer (infra vendor)  FFS whether coordination with encoder vendor is required | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on the RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | Type 1 and 3 (note 1)  Up to DUT vendor | Type 2 and type 3 with NW-first training (note 2)  Up to coordination between DUT vendor and decoder vendor | Type 3 (note 3)  Up to DUT vendor | Type 2 and type 3 with NW-first training (note 2)  Up to coordination between DUT vendor and TE vendor |
| Test decoder performance verification procedure at TE ~~and/or DUT~~ | – need to ensure that decoder performance is not degraded(as intended by the decoder provider) on the TE | - need to ensure that decoder performance is not degraded(as intended by the decoder provider) on the TE  – need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long as the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS  Yes (e.g., DUT can provide some sample dataset for reference) | FFS  Yes for 1st bullet in above row (e.g., decoder vendor can provide some sample dataset and encoder for reference)  Not clear for 2nd bullet in above row. This seems to be contradictory with the goal of test. | FFS  No need to discuss | FFS  No need to discuss |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used in actual field deployments )~~knowledge of model, training type, etc.)~~ | Low, may not reflect the actual decoder implemented by infra vendor | High | Medium/low, depends on training data used when decoder is specified | Low/Medium  depends on training data used when decoder is specified |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower than Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/interoperability between TE and DUT needs further study. |
| Specification Effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs study and decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure(after specs are published) | Model exposure is required from DUT to TE vendor.  FFS detailed approach to share the decoder. | Model exposure is required from decoder vendor to TE vendor and DUT (under training type 3).  FFS detailed approach to share the decoder. | No | No |
| Applicability to different scenarios/conditions/ configurations | Yes (note 4) | Yes | Yes | Yes |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendors and DUT vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than option 1. | Testing the encoder at DUT  Low –providing no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – providing no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than option 3/4  FFS compared to option 2 | Higher than Option 3/4  FFS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | High   * The gap between the test decoder from DUT and actual deployed decoder from infra vendor. * Need for interaction between DUT and infra vendor | Low | Low | Lower than option 1   * The gap between the test decoder from TE vendor and actual deployed decoder from infra vendor * Need for interaction between TE and infra vendor |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | High | High | Low | Low/medium |
| Relationship with reference decoder/encoder(used by RAN4 to define the performance requirements) for defining requirement | No direct relationship. | No direct relationship. | Reference decoder are same as test decoder | May or may not be the same.  Reference decoder could be developed based on the same parameters/conditions that to be specified for defining test decoder. |
| Whether model transfer/delivery is needed during the test procedure | DUT needs to pass the decoder to TE. This could be a step before test.  No need during the performance test procedure.  FFS LCM test. | Decoder vendor needs to pass the decoder to TE. This could be a step before test.  No need during the performance test procedure.  FFS LCM test | No need | No need |

R4-2400562:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Option 1: DUT provides decoder** | | **Option 2: Decoder not from DUT and Spec** | **Option 3: Full decoder specification in standard** | **Option 4: TE specified decoder** |
| Clarification of options | | | | | |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | Defer to WI | | | | |
| Pros/Cons analysis | | | | | |
| Reflection on the real deployment   * Knowledge of the test decoder * Whether the decoder vendor can implement the test decoder | * Not reflected * Depends on what is specified | | * Is reflected * Can implement within the same vendor | * Not reflected * Can implement | * Is reflected * Depends on what is specified |
| Confidentiality/ IP issues | Yes | | Yes | No | No |
| Applicability to different scenarios/conditions/ configurations | No difference across scenarios | | | | |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Depends on agreed high level parameters and whether/how to ensure test repeatability and the ability to implement decoders with similar performance by other vendors | | | May not have forward compatibility | Depends on agreed high level parameters and how to ensure test repeatability and the ability to implement decoders with similar performance by other vendors |
| Relationship with reference decoder/encoder for defining requirement | For simulation alignment purpose | Same as option 4 | | For simulation alignment purpose | For dataset generation, decoder verification and simulation alignment purpose |

R4-2401045:

**Table 7.4.2.3-1 Comparison of the four options of test decoder (TR 38.843)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendor (no need to be specified) | Up to decoder implementer (infra vendor)  FFS whether coordination with encoder vendor is required | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on the RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) |  |  |  |  |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | - Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  - Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long as the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used in actual field deployments ) |  |  |  |  |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower than Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/interoperability between TE and DUT needs further study. |
| Specification Effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs study and decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure(after specs are published) |  |  | No | No |
| Applicability to different scenarios/conditions/ configurations |  |  |  |  |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than option 1. | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than option 3/4  FFS compared to option 2 | Higher than Option 3/4  FFS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem |  |  |  |  |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented |  |  |  |  |
| Relationship with reference decoder/encoder(used by RAN4 to define the performance requirements) for defining requirement |  |  |  |  |
| Whether model transfer/delivery is needed during the test procedure |  |  |  |  |

R4-2401172

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendor (no need to be specified) | Coordination with encoder vendor is required | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | Alignment of training data between UE and TE may be required. |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on the RAN4 specification |
| Supported training collaboration type (source of training data should be consistent with the collaboration type) | FFS | FFS | FFS | FFS |
| Test decoder verification procedure at TE and/or DUT | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | - Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  - Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long as the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | Yes. | FFS.  If the performance of decoder will be verified, what’s the reference encoder? If different encoders from different UE vendor are provided, how to verify the decoder?  The decoder verification seems to be encoder specific. | Yes. | FFS.  Similar as option 2. |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (knowledge of model, training type, etc.) | No, there may be mismatch between decoder from UE and NW vendor | Yes. | No, there may be mismatch between decoder from specification and NW vendor | Depends on what’s partially specified for the decoder. There may be mismatch between decoder from specification and TE |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE.  Lower than Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE.  Lower than Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower than Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/interoperability between TE and DUT needs further study. |
| Specification Effort (e.g. test decoder) | Low | Low | Highest   * RAN4 needs to standardize the entire decoder | High  RAN4 needs study and decide on what to standardize |
| Confidentiality/IP issues | Need to be considered | Need to be considered | No. | No. |
| Applicability to different scenarios/conditions/ configurations | Depends on how to design the test to guarantee the generalization | Depends on how to design the test to guarantee the generalization | Depends on how to design the test to guarantee the generalization | Depends on how to design the test to guarantee the generalization |
| Complexity of actual testing procedure for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than option 1. | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than option 3/4  FFS compared to option 2 | Higher than Option 3/4  FFS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem |  |  |  |  |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes. | No. depends on the implementation of decoder in NW vendor.  Offline alignment may still be needed. | Depends.   * If new AI decoder is invented which requires simpler encoder. With simpler new encoder, UE may not pass the test with old decoder. * If new AI encoder is invented which requires more simple decoder. With more advanced encoder, UE can pass the test with old decoder. | No. depends on the implementation of decoder in TE vendor.  Offline alignment may still be needed. |
| Relationship with reference decoder/encoder for defining requirement | If reference decoder is defined, a reference encoder maybe needed either. Otherwise, how to verify the performance of reference decoder when there are many encoders is FFS. | If reference decoder is defined, a reference encoder maybe needed either. Otherwise, how to verify the performance of reference decoder when there are many encoders is FFS. | Reference encoder is not needed. | Depends on what’s specified for the decoder. |
| Whether model transfer/delivery is needed during the test procedure | Not specific to two side model. | Not specific to two side model. | Not specific to two side model. | Not specific to two side model. |

R4-2401612:

Table 2.4-1 Comparison of the four options of test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendors (no need to be specified) | * Up to decoder implementer (infra vendor) | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS   * Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on RAN4 specification |
| ~~Supported training~~ ~~collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type)~~ | ~~Up to DUT vendor (All training collaboration Type 1/2/3)~~ | ~~Up to infra vendor (All training collaboration Type 1/2/3)~~ | ~~Up to RAN4 procedure to specify the decoder~~ | ~~Up to TE vendor (All training collaboration Type 1/2/3)~~ |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long a the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used | Low  There could be large performance mismatch with field performance due to mismatch between test decoder and field decoder implemented by infra vendors.  Depends on training collaboration type and/or training dataset, the decoder mismatch would be alleviated. | Medium  Could reflect the performance in the field if network vendors use same or similar decoder in the field as the test decoder.  Since test decoder is designed for minimum requirement, network vendors may use more powerful decoder with better performance in the field. | Low/Medium  Could reflect the performance if the test decoder(s) is generated from the well-designed datasets that could reflect real deployment.  There could be large performance mismatch if the training dataset is not realistic. UE may have to implement an additional encoder only for the tests. | Medium  Could reflect the performance if the test decoder(s) is generated from the well-designed datasets that could reflect real deployment.  Could reflect the performance if infra/UE vendors consider the partially specified test decoder as reference for implementation. |
| TE requirements to deploy the decoder (e.g., training, complexity, interopereatbility) | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower thank Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/ interoperability between TE and DUT needs further study |
| Specification effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs to study and may decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure (after specs are published) | Yes  DUT vendor might have to expose some aspects of the design to the TE vendor  Depending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing | Yes  Decoder vendor might have to expose some aspects of the design to the TE vendor  Depending on means used to share test decoder, TE vendors might require integrating source code from third party, which could even require licensing | No | No |
| Applicability to different scenarios/conditions/ configurations | Applicable  Depending on how generalization test is defined and how test decoder is trained. | Applicable  Depending on how generalization test is defined and how test decoder is trained. | Applicable  Depending on how generalization test is defined and how test decoder is trained. | Applicable  Depending on how generalization test is defined and how test decoder is trained. |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than Option 1 | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than Option 3/4  FSS compared to Option 2 | Higher than Option 3/4  FSS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | High  Offline co-engineering between TE vendor and UE vendors may be needed depends on model format.  TE needs to select different test decoder for different DUT, which may be based on DUT declaration.  All UE vendors should develop its own test decoder. | High  Offline co-engineering between TE vendor and infra vendors may be needed depends on model format.  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?  Whether should all infra vendors provide test decoder?  DUT may need to be tested against one or multiple test decoders provided by different infra vendors. | Low  TE only needs to implement the test decoder.  DUT may consider the test decoder for encoder implementation | Low/Medium  TE only needs to train and implement partially specified test decoder.  DUT may consider the test decoder for encoder implementation |
| Friendly to STOA (state of the art) model test / Forward compatibility when new AI models are invented | Yes | Yes | No | Yes |
| Relationship with reference decoder/encoder (used by RAN4 to define the performance requirements) for defining the requirement | A different reference decoder (e.g., based on option 3 or option 4) for defining requirements. | A different reference decoder (e.g., based on option 3 or option 4) for defining requirements. | Same reference decoder as test decoder for defining requirements. | Same reference decoder as test decoder for defining requirements. |
| Whether model transfer/delivery is needed during the test procedure | FFS | FFS | FFS | FFS |

R4-2401687:

Table 1: Comparison of the four options of test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Option 1 | Option 2 | Option 3 | Option 4 |
| Clarification of options | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendors (no need to be specified) | Up to decoder implementer (infra vendor) | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined   * Option 4a: If model structure is not specified, a sample-by sample training data should be specified, where each sample consists of both raw channel/precoding matrix and the bit stream of CSI feedback. * Option 4b: If model structure is specified, the consistency between training data and test data should be ensured. Whether to align sample-by-sample training dataset depends on the divergency of simulation results among companies. |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | Wait for training collaboration types that are specified in other WGs. | | | |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| Pros/Cons analysis | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used | Wait for training collaboration types that are specified in other WGs. | | | |
| TE requirements to deploy the decoder (e.g., training, complexity, interopereatbility) | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower thank Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/ interoperability between TE and DUT needs further study |
| Specification effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs to study and may decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure (after specs are published) | YES | YES | No | No |
| Applicability to different scenarios/conditions/ configurations | YES | YES | YES | YES |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than Option 1 | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than Option 3/4  FSS compared to Option 2 | Higher than Option 3/4  FSS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | High | High | Low | Low |
| Friendly to STOA (state of the art) model test / Forward compatibility when new AI models are invented | Not applicable | Not applicable | Not applicable | Not applicable |
| Relationship with reference decoder/encoder (used by RAN4 to define the performance requirements) for defining the requirement | Test decoder is NOT the reference decoder. | Test decoder is NOT the reference decoder. | Test decoder is the reference decoder. | Test decoder is the reference decoder. |
| Whether model transfer/delivery is needed during the test procedure | YES | YES | NO | NO |

R4-2401817:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendor (no need to be specified) | Up to decoder implementer (infra vendor)  FFS whether coordination with encoder vendor is required | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on the RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | More suitable for type1 UE side training and type3 UE first training.  With these training collaboration types (type1 UE side and type3 UE first), in practical, it is UE to prepare CSI encoder and corresponding CSI decoder first, and then provide CSI decoder (or CSI decoder data) to NW for use. Therefore, it is more reasonable for UE to also provide the test decoder in RAN4 tests. | More suitable for type1 NW side training and type3 NW first training.  With these training collaboration types (type1 NW side and type3 NW first), in practical, it is NW to prepare CSI encoder and corresponding CSI decoder first, and then provide CSI encoder (or CSI encoder data) to UE for use. Therefore, it is more reasonable for NW to also provide the test decoder in RAN4 tests. | Support all  The way a model is trained can be decoupled from how the model is tested. | Support all  The way a model is trained can be decoupled from how the model is tested. |
| **In general, the way a model is trained can be decoupled from how the model is tested.**  However, for certain training collaboration types, testing methods can be better matched to deployment/utilization scenarios.  For example, for option1 and option2,  (1) with type1 UE side training(UE trains a encoder and a decoder, then UE transmits the decoder to NW), it is more reasonable to use option1 in RAN4 tests, where the UE provides the test decoder, because in practical use, the UE also provides the decoder.  (2) with type1 NW side training(NW trains a encoder and a decoder, then NW transmits the encoder to UE), it is more reasonable to use option2 in RAN4 tests, where the NW provides the test decoder, because in practical use, the NW also have the decoder.  Similarly,  (3) with type3 UE first training(UE trains a encoder and a decoder, then UE transmits a data set to NW side, which can be used for decoder model training), it is more practical to use option1 in RAN4 testing,  (4) with type3 NW first training(NW trains a encoder and a decoder, then NW transmits a data set to UE side, which can be used for encoder model training), it is more practical to use option2 in RAN4 testing.  **The impact of different training collaboration types on the provider of encoder and decoder in practical use should be considered, and evaluate whether different testing options can be utilized accordingly to better align with actual use cases under such impact.** | | | |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | - Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  - Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long as the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used in actual field deployments ) | Low,  Lower than option2,  Depends on training data | Low, depends on training data  As a data/scenario driven solution, AI/ML models be utilized in different cells may differ from each other.  A limited number of test models, even if provided by NW, cannot fully reflect the real deployment scenarios(e.g. different cells/scenarios/ channel conditions) | Low,  Lower than option1 and option 2,  Depends on training data | Low,  Lower than option1 and option 2,  Depends on training data |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder are implemented by TE  Lower than Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower than Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/interoperability between TE and DUT needs further study. |
| Specification Effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs study and decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure(after specs are published) | TBD,  Confidentiality/ IP issues should be case/model/solution specific, hard to have a general conclusion | TBD  Confidentiality/ IP issues should be case/model/solution specific, hard to have a general conclusion | No | No |
| Applicability to different scenarios/conditions/ configurations | Low,  similar to the question “Reflection on the real deployment” | Low  similar to the question “Reflection on the real deployment” | Low  similar to the question “Reflection on the real deployment” | Low  similar to the question “Reflection on the real deployment” |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than option 1. | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than option 3/4  FFS compared to option 2 | Higher than Option 3/4  FFS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | Medium | High  May involve too many entities, such as UE, NW, and TE, in a single test | low | low |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes | Yes | No | [No] |
| Relationship with reference decoder/encoder(used by RAN4 to define the performance requirements) for defining requirement | May need a separate reference decoder/encode | May need a separate reference decoder/encode | May reuse test encoder/decoder as reference encoder/decoder | May reuse test encoder/decoder as reference encoder/decoder |
| Whether model transfer/delivery is needed during the test procedure | Yes | Yes | No | No |

R4-2402390

**Proposal 2: The following clarification of options are provided for option 1-4 test decoder for 2-sided model (except the green highlighted part for existing agreement).**

Table 1: Comparison of the four options of test decoder

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Option 1** | **Option 2** | **Option 3** | **Option 4** |
| **Clarification of options** | | | | |
| Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specifications | TE vendor, decoder developed based on RAN4 specifications |
| Source of decoder training data | Up to DUT vendors (no need to be specified) | Up to decoder implementer (infra vendor) | Not needed, decoder fully specified (used as part of the RAN4 procedure to specify the decoder) | ~~FFS~~  ~~Could be specified depending on how Option 4 will be defined~~  Training data set, specified in 3GPP standard (Proposal 1) |
| DUT vendor knowledge of the test decoder | Full knowledge | No or partial or enough or full knowledge based on alignment with infra vendors or specifications | Full knowledge based on the specifications | Partial knowledge – based on RAN4 specification |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | **Type 1**  (Joint training of encoder/decoder  at UE-sided) | **Type 2 or Type 3 (NW first)**  (offline collaboration available between UE and gNB vendors is required) | **No collaboration required** (test decoder is standardized and open to DUT vendors) | **No collaboration required, and procedure is** **similar to Type 3 (NW first)** (the dataset is standardized already and actually no collaboration between TE and DUT vendors needed) |
| Test decoder performance verification procedure at TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE | Need to ensure that decoder performance is not degraded (as intended by the decoder provider) on the TE  Need to ensure that decoder performance is good enough to enable a DUT that meets the minimum requirements to pass the test | Not needed as long as the standardized model implementation can be similar enough between TE vendors | Not needed as long a the model implementation can be similar enough between TE vendors |
| Feasibility of test decoder verification procedure | ~~FFS~~ **Procedure needs to be clarified** (During this verification in particular condition, performance shall be guaranteed based on a reference encoder also provided by decoder vendor) | ~~FFS~~ **Procedure needs to be clarified** (During this verification in particular condition, performance shall be guaranteed based on a reference encoder also provided by decoder vendor) | ~~FFS~~ Not applicable | ~~FFS~~ Not applicable |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (likelihood that test decoder would be used | **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 relevant collaboration is available in the real deployment) | **Maybe** (Depends on whether specified test decoder can reflect decoder in the field) | **Maybe** (Depends on whether specified data set for training can reflect decoder in the field) |
| TE requirements to deploy the decoder (e.g., training, complexity, interopereatbility) | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Higher than Option 3/4 in terms of that maybe more than one decoder is implemented by TE  Lower thank Option 3/4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Lower thank Option 4 in terms of that no training at TE is required | Lower complexity than Option 1/2 in terms of that only one decoder is implemented by TE  Higher than Option 3 in terms of that training at TE is required  Note: How to ensure compatibility/ interoperability between TE and DUT needs further study |
| Specification effort (defining test decoder and requirements) | Low | Low | Highest  RAN4 needs to standardize the entire decoder | High  RAN4 needs to study and may decide on what to standardize |
| Confidentiality/ IP issues in the testing procedure (after specs are published) | **Yes** (Disclosure of UE vendor designed IP during testing) | **Yes** (Disclosure of BS  vendor designed IP during testing) | No | No |
| 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 training data set for different scenarios/conditions /configurations  accordingly |
| Complexity of testing for the ecosystem | Testing the encoder at DUT  Higher than Option 3/4  Need for interaction between TE vendor | Testing the encoder at DUT  Higher than Option 3/4  Testing complexity higher also than Option 1 | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties | Testing the encoder at DUT  Low – no need for interaction between TE vendors and other parties |
| Complexity of verifying/testing the test decoder | Higher than Option 3/4  FSS compared to Option 2 | Higher than Option 3/4  FSS compared to Option 1 | Low | Low |
| Complexity of deploying for the ecosystem | Not sure which is different from the row of “Complexity of testing for the ecosystem”, propose to remove this row. | | | |
| Friendly to STOA (state of the art) model test / Forward compatibility when new AI models are invented | **Friendly to SOTA**, but depends on gNB can adopt the newly developed decoder by UE in practice | **Friendly to SOTA**, as long as new AI model (for encoder part) is tested with gNB developed decoder before pushing to UE | **Friendly to SOTA**, as long as new AI model (for encoder part) is tested with standardized reference decoder before pushing to UE | **Friendly to SOTA**, as long as new AI model (for encoder part) is tested with TE developed decoder before pushing to UE |
| Relationship with reference decoder/encoder (used by RAN4 to define the performance requirements) for defining the requirement | Reference decoder/encoder for defining the requirement needs separate discussion in RAN4 | Reference decoder/encoder for defining the requirement needs separate discussion in RAN4 | Encoder can be developed individually be vendors for performance alignment in RAN4 | Encoder can be developed individually be vendors for performance alignment in RAN4 |
| Whether model transfer/delivery is needed during the test procedure | **No** (Test decoder is provided by UE vendors before the test procedure) | **No** (Test decoder is provided by gNB vendors before the test procedure) | **No** | **No** |