**3GPP TSG-RAN WG1 Meeting #109 R4-2318142**

**Chicago, USA, November 13th – 17th 2023**

**Agenda item:** 8.21.4

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [109][136] FS\_NR\_AIML\_air

**Document for:** Information

# Introduction

This is the summary thread for issues related to NR AI/ML study in RAN4. A WF summarizing many topics/issues to be further studied and discussed was agreed in the previous meeting in R4-2317258. This summary is organized in 3 high level topics and contains several sub-topics for discussion.

# Topic #1: General aspects and TR

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2318250**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318250.zip) | CAICT | **Proposal 1: Support option 3 as the testing goals (issues 1-4) for AI/ML related testing.**  **Proposal 2: Latency requirements of data collection for model inference and monitoring could be considered per use case and further discussed in WI.**  **Proposal 3: Suggest to approve the tentative agreement in Issue 2-5 (Accuracy requirements for measurement data or label data).**  **Proposal 4:** **Explicit definition of ground truth could be discussed further in WI after necessity being identified with consensus for each use case.** |
| [**R4-2318281**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318281.zip) | CATT | **Observation 1: Conditions/additional conditions are use case-specific and will be discussed by RAN1 in WI phase.**  **Proposal 1: RAN4 waits for RAN1 progress on configurations and discuss details of test scenarios first.**  **Proposal 2: Minimum level performances are scenario/configuration-specific and can be discussed in WI phase.**  **Proposal 3: Performance degradation should be discussed case by case. And different approaches may be able to assess the degradation in one scenario/use case.**  **Proposal 4: Whether option 2, i.e., RAN4 defines one test and changing different propagation conditions within the test, is used to verify the generalization ability in tests depends on TE implementation.**  **Proposal 5: Regarding the testing goal, option 3 (option 1 and option 2 depending on the test) is preferred.**  **Proposal 6: R****AN4 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.**  **Proposal 7: RAN4 to discuss the ground truth case by case since it is related to reference point selection which varies for different types of equipment and FR.** |
| [**R4-2318478**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318478.zip) | NTT DOCOMO, INC. | **Proposal 1: About generalization performance verification, the minimum level performance which has to be achieved/maintained under the identified scenario and/or configuration should be same as existing requirements if it is the use case with the existing legacy performance.**  **Proposal 2: There are possibly two directions to define significant degradation for other scenarios and/or configurations:**   * **Taking typical validation methods which are hold-out validation, K-fold cross validation, etc**   + **The dataset for validation and performance metric should be specified in the spec** * **Taking the relative performance validation**   + **If the objective use case can be operated without AI, the significant degradation can be defined as "lower than without AI performance"** |
| [**R4-2318489**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318489.zip) | CAICT, Qualcomm, Ericsson | Text proposal for TR 38.843 capturing the RAN4 part based on agreements reached so far. |
| [**R4-2318579**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318579.zip) | Apple | **Proposal 1:** **The post deployment testing should be based on the model monitoring framework**  **Proposal 2: Based on status RAN1 study on model monitoring in Rel 18, the post deployment testing feasibility study should be postponed to Rel 19 as a part of a study phase. Feasibility study shall include the discussion of defining requirements to ensure the accuracy of reporting reliability metrics is consistent among different UEs to allow the NW to manage the AI functionality.**  **Proposal 3: If the feasibility is confirmed, RAN4 could further define a testing framework to enable model monitoring, performance degradation validation** |
| [**R4-2318935**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318935.zip) | Qualcomm, Inc. | **Proposal 1: RAN4 testing reference block diagram is based on the agreements captured in the WFs from previous meetings.**  **Proposal 2: RAN4 testing reference block diagram is based on the following agreements**   * **Model inference (R4-2306299)**   ***…. AI/ML based performance enhancements mainly focus on how to define requirements and tests for inference***   * **Functionality/Model(M/F) monitoring procedure (R4-2306299)**   **(Study) *Performance (Model/Functionality) monitoring procedure, including performance evaluation and decision-making procedure …***   * **Functionality/Model(M/F) management procedure (R4-2306299)**   **(Study) *Functionality/Model management procedure, including functionality/model selection/activation/deactivation, and functionality/model switching/fallback/transfer/delivery/update***   * **Test data generator (R4-2306299)**   **(Study) *Different generating methods of test dataset can be used for different tests. The following candidate methods are to be considered …***  **Proposal 3: One-sided model block diagram: note that the data collection is pending other group decision and FFS in the WF, therefore we present it in dashed box, and companies can comment to decide whether to keep it.**  Data collection  M/F management  Verification  AI/ML modules  AI/ML modules  Test configuration/controller  Test data generator  Model inference  M/F management  M/F monitoring  DUT  TE  **Proposal 4: Two-sided model block diagram: same note as proposal 3.**  Data collection  M/F management  Verification  AI/ML modules  AI/ML modules  Test configuration/controller  Test data generator  Model inference  M/F management  Model inference  M/F monitoring  DUT  TE |
| [**R4-2319075**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319075.zip) | vivo | ***Observation 1: The AI/ML complexity could be considered when specify reference model for defining requirements and discussed case by case, e.g. model size, FLOPS.***  ***Observation 2: Post deployment performance may be verified by model monitoring.***  ***Observation 3: “Ground truth” in RAN4 need to be discussed on a use case by use case basis.***  ***Proposal 1: Accuracy requirements for input data collection need to be considered. (at least for positioning)***  ***Proposal 2:*** ***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 3: RAN4 testing goal for AI/ML is to verify whether a specific AI/ML model can be conducted in a proper way and whether the performance gain of AI/ML model can be achieved for specified scenario/configuration.***   * ***Define reference models for different use cases with corresponding requirements as the specific AI/ML model for verification.*** * ***Static channel models (TR 38.901) based dataset used in legacy RAN4 test can be used as the baseline as the specific scenarios/configurations.*** * ***The specific testing goal depending on the test.***   ***Proposal 4: The feasibility of generalization test by using dynamic scenarios/configurations needs further study by considering the following aspects.***   * ***Testing method***   + ***TE changes the scenario/channel in turn and covers all the required N scenarios/channels to be tested***   + ***TE chooses a subset of the N scenarios/channels as the scenarios/channels under test based on the certain rules, and changes the scenario/channel randomly and completes the test within the requirement time T.*** * ***Test dataset generation***   + ***Stationary statistical channel modelling method***   + ***Non-stationary channel modelling method***   + ***Field channel measurement***   + ***Deterministic channel modelling (e.g., Ray Tracing)*** |
| [**R4-2319644**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319644.zip) | Ericsson | For Rel-19 work, RAN is still discussing what will start as normative work and what will or will not have further study in RAN1. From a RAN4 perspective, even if topics become more mature in RAN1, it is important that sufficient progress is made in developing an understanding of RAN4 requirements and RAN4/5 tests that RAN4 is in a good position to develop specifications. To enable this, in our understanding RAN4 discussions need to in some way continue into 2024. Thus, we propose that RAN4 recommend that RAN4 discussion should continue/begin right from the start of any WI.   1. RAN4 should recommend to continue/start discussion from the beginning of any AI-PHY WI   It is also important that RAN provides sufficient time units to adequately cover all of the issues listed above. The amount of work is quite significant and it is important to have sufficient meeting time to handle all of the issues.  A quite significant amount of TU should be allocated for AI/ML (exact amount depending on the scope and number of use cases; for all use cases 2-3 TU should be considered). |
| [**R4-2319824**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319824.zip) | Nokia, Nokia Shanghai Bell | 1. For all RAN4 issues and discussions, it is important for RAN4 to seek alignment with current RAN1 and RAN2 discussions and agreements regarding AI/ML functionality and model ID – based LCM. 2. For issue 1-2 (Generalization goals), RAN4 to adopt as the way forward previous agreement with the following modifications:  * Verify whether the ~~performance gain/~~minimum level of performance of AI/ML functionality~~/model~~ can be achieved/maintain under various propagation conditions ~~scenarios~~ and~~/or~~ functionality configurations (based on UE capabilities), while the performance won’t be significantly degraded in other propagation conditions ~~scenarios and/or configurations~~.  1. For the second and third FFS in issue 1-2 (Generalization goals), RAN4 to agree that these are use case specific aspects which need to be determined based on the RAN1 studies during the SI/WI complemented with RAN4 specific simulations during the WI. 2. For issue 1-3 (Handling of generalization in tests), RAN4 to adopt as way forward the option 1 with the following modifications:  * **Modified Option 1: Signaling for Functionality-basedLCM procedures and functionality performance monitoring are considered in dedicated test cases and are excluded in tests verifying generalization and robustness. RAN4 defines multiple tests with different propagation conditions. In each of the test, TE can configure~~s~~ the same model ID when available from the UE, and therefore the same AI/ML Functionality ~~model~~ is tested under different propagation conditions to verify it’s generalizability and robustness. (environment differs in each test but not changing dynamically during the test).** * **Specified UE configuration includes functionality and~~/or~~ model ID ~~if defined~~ when available from the UE**  1. For ‘FFS on Option 2’ in issue 1-3 (Handling of generalization in tests), RAN4 to adopt as way forward the following modification:  * **In Option 2, change the same model ID to “the same specified UE configuration, which includes functionality and~~/or~~ model ID ~~if defined~~ when available from the UE.**  1. RAN4 requirements and test procedures should be defined on the level of ML-enabled Functionality/Feature, i.e., model- specific requirements and tests shall be precluded. 2. For Issue 1-4 (Testing goals), RAN4 to adopt as way forward the option 2a with the following modifications:  * **Option 2a: The testing goal is to verify whether the minimum performance~~/performance gain~~ of AI/ML ~~model/~~functionality/feature can be achieved for a static ~~or~~ and non-static(dynamic) scenario/configuration.**    + - **FFS how to define the static test scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)**     - **FFS how to define the minimum performance target(s) (e.g., by defining AI/ML dedicated performance/core requirements associated with use cases)**     - **FFS how to define the non-static specific scenarios/configurations**  1. For Issue 1-5 (Latency requirements) proposals, RAN4 to adopt the following modified text (based on the initial issue 1-5):    * **Option 1: Consider data collection latency requirements only for inference and monitoring**    * **Option 2: RAN4 shall define the latency requirements based on RAN2’s agreements and the MAX total latency requirements can be:**    * **Latency requirements for functionality performance monitoring data collection**      + **Option 3.1: Latency requirements of data collection for ~~model inference and~~ functionality performance monitoring should be considered and discussed per use case, subject to the output from RAN1/2.**      + **Option 3.2: RAN4 should study latency requirements for data collection of ~~model~~ functionality performance monitoring, at least for positioning and CSI compression use cases.**    * **Latency requirements for inference data collection**      + **Option 4.1: Latency requirements of data collection for model inference ~~and monitoring~~ should be considered and discussed per use case, subject to the output from RAN1/2.**    * **Latency requirements for training data collection**      + **Option 5.1: Do not study latency requirements for training data collection, discuss latency requirements for any particular use case during WI as needed** 2. For the *modified* Issue 1-5 (Latency requirements) proposals, RAN4 to adopt Option 1 and 2 as baseline, and Option 3.1 and Option 4.1 as way forward. 3. For issue 1-8, RAN4 should further study how the mutual impact of several simultaneously supported and activated ML-enabled Functionalities/Features to ensure the absence of performance degradation. Details of added testing features and evaluation procedure can be left to the work item phase. 4. The recommended WF for “**Issue 1-9: Post deployment testing**” is *Option 2: RAN4 does not need to study such framework*. However, the initial conformance testing of the AI/ML functionality cannot ensure the same level of performance for the devices in the field. And for the same reason, post deployment validation of the functionality becomes an issues. 5. 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:    1. The changes/updates to the ML-enabled Functionality/Feature are tested and declared by the device vendor against RAN4 requirements before any deployment to the UE is performed.    2. After deployment to the UE and before changed/updated ML-enabled Functionality/Feature is activated in the UE, a post-deployment validation is performed, e.g., a sanity check test loop is run, e.g., using the functionality performance monitoring and LCM activation/deactivation/switching procedures,    3. At least one fallback/default Functionality/Feature that passed conformance testing must always be present in the device. |
| [**R4-2319825**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319825.zip) | Nokia, Nokia Shanghai Bell | Text proposal with the test diagrams |
| [**R4-2320183**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320183.zip) | Huawei,HiSilicon | TP for 38.843 based on the Ran4 agreements in the previous meetings |
| [**R4-2320184**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320184.zip) | Huawei,HiSilicon | ***Observation 1:*** 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 1:*** Whether to take legacy baseline requirements also as baseline requirements for AI/ML-specific feature should be discussed per use case.   * Note: In some cases, legacy test metrics are not testable at least based on RAN4 current studying.   ***Proposal 2:*** RAN4 AI/ML testing goal is identified from the following options.   * Option 1: The testing goal is to verify whether a specific AI/ML model can be conducted in a proper way.   + FFS how to determine the specific AI/ML model.   + FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs) * Option 2: The testing goal is to verify whether the performance gain of AI/ML model can be achieved for a specific scenario/configuration.   + FFS how to determine the specific scenario/configuration, taking account of ensuring the consistency between testing dataset and training dataset.   ***Proposal 3:*** Table for RAN4 testing goal when model under test is transferred from the opposite side with and w/o air interface signaling.   |  |  |  | | --- | --- | --- | |  | Testing goal (Option 1): verify that the model is properly conducted | Testing goal (Option 2): verify the performance of the model | | Model under test is transferred from the opposite side with air-interface signaling | √ | - | | Model under test is delivered w/o air interface signaling | - | √ |   Note1: Wait RAN1/2 to study whether to specify model transfer with air interface signaling or not.  Note2: Whether it is testable or not for selected testing goal is a separate discussion.  ***Proposal 4:*** Whether to consider generalization verification needs to wait RAN1/2 progress.   * If model transfer over the air interface signaling is not specified and generalization is testable after RAN4 studying, then generalization verification reuses legacy RAN4 test, where different requirements may be considered in different scenario/configuration, separately.   ***Proposal 5:*** Take functionality-based LCM as the starting point for RAN4 discussion.  ***Proposal 6:*** Identified scenarios and/or configurations can be initially interpreted as the scenarios and/or configurations that UE report by capability signaling.  ***Observation 2:*** A large range of various UE capabilities may be involved, which is problematic for RAN4 to specify the test cases.  ***Proposal 7:*** 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 8:*** 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.  Note: There is no consensus from other WGs on whether and how to define UE supported site-specific configuration/channel conditions in UE capability.  ***Observation 3:*** If legacy test metrics are not valid/testable when defining AI/ML-specific requirements, legacy performance requirements for non-AI cannot be reused.  ***Proposal 9:*** RAN4 will study the minimum level performance, per use case, for identified scenarios and/or configurations (if specified).  ***Proposal 10:*** 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 4:*** There is no need to introduce AI/ML-related requirements in the other scenarios and/or configurations.  ***Proposal 11:*** Performance for other scenarios and/or configurations can be ensured by RAN4 legacy test.  ***Observation 5:*** There is no benefit identified by by introducing channel condition changes during test.  ***Proposal 12:*** RAN4 not considers introducing channel condition changes during test. |
| [**R4-2320357**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320357.zip) | Ericsson | TP with update to terminology  **Test encoder/decoder for TE**: AI/ML model for UE encoder/gNB decoder implemented by TE. |
| [**R4-2320554**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320554.zip) | ZTE Corporation | **Observation 1: RAN2 assumes that there is no latency requirements for data collection for all types of offline training. However, the latency requirements shall be studied for data collection for inference data and monitoring data.**  **Observation 2: For model inference, the MAX total latency values will be .**  **Proposal 1: RAN4 shall define the latency requirements based on RAN2’s agreements and the MAX total latency requirements can be:**    where,  **t1** denotes the propagation delay from gNB to send the related RS to UE;  **t2** denotes the measurement time for RS;   * **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.  **In theory, t1=t4 the propagation delay shall be the same which equals the distance (d) between gNB and UE divides the speed of light (c):**    **Proposal 2: The latency requirements shall be defined per use case**.  **Observation 3: The ground truth shall be mapped to the input and compared to measurement results to guarantee the data quality.**  **Observation 4: If model is sent to the field which the radio conditions and other obstacles ( e.g. NLOS) are totally different and complicated than the test environment, the model performance will be degradation and also the model generalization is not beneficial to RAN4 testing goals.**  **Proposal 3: The scenario of UE capability report, which is AI authentication, it can be identified. Otherwise, the remaining scenarios can be regarded as the other scenarios for test.**  **Proposal 4: The minimum level of performance shall be studied per use case.**  **Proposal 5: The legacy performance under different use cases can be as the baseline to judge the minimum performance of AI-based method for each identified scenario.**  **Proposal 6: The margin shall be added based on the known performance gain for identified scenario to judge what is the significant degradation for other scenarios and try to guarantee that it will not occur**. |
| [**R4-2320610**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320610.zip) | Samsung | *Handling of generalization*  **Proposal 1: Further understanding on generalization goals is provided/highlighted:**   * Verify whether the performance gain/minimum level of performance of AI/ML functionality/model can be achieved/maintain under the identified scenarios and/or configurations, while the performance won’t be significantly degraded in other scenarios and/or configurations   + ~~FFS on details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality~~     - The identified scenarios and/or configurations for test       * shall include: Configuration(s) of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature/FG, for AI/ML model/functionality identification and model-ID-based/functionality-based LCM of UE-side models and/or UE-part of two-sided models;       * shall NOT include: Additional conditions (e.g., any aspects that are assumed for the training of the model but are not a part of UE capability for the AI/ML-enabled feature/FG) as determined/identified between UE-side and NW-side   **Proposal 2: The following two bullets shall be discussed in normative phase by focusing on the particular use case, and no further discussion is needed in SI:**   * **FFS on what the minimum level performance for each identified scenario and/or configuration is** * **FFS on what the significant degradation for other scenarios and/or configurations is**   **Proposal 3: For handling of generalization in tests, RAN4 adopt modified Option 1.**  *Requirement for data collection*  **Proposal 4: For data collection for training purpose:**  **- Latency requirement shall not be specified in RAN4;**  **- The importance of defining accuracy requirement is use case-specific: Until a clear data collection procedure is defined in RAN1/2 (, i.e., probably in work item stage), there is no necessity for RAN4 to further study on accuracy requirement for data collection for training purpose.**  **Proposal 5: For data collection for inference purpose,**  **- RAN4 requirement shall be specified only if the required data comes from other entities and the procedure of data collection is agreed in RAN1/2.**  **- The necessity of latency and/or accuracy requirement should be discussed for the particular use case for normative work.**  **Proposal 6: the data collection for monitoring purpose,**  **- RAN4 requirement shall be specified only if the required data comes from other entities and the procedure of data collection is agreed in RAN1/2.**  **- The necessity of latency and/or accuracy requirement should be discussed for the particular use case for normative work.**  *Requirements for LCM*  **Proposal 7: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case for normative work.**  **- No need further RAN4 study until clear monitoring metrics/methods specified in other WGs.**  *RAN4 performance testing goals*  **Proposal 8: For RAN4 performance testing goal, Option 3 (Option 1 or 2 depending on test) is preferred to guarantee LCM procedure and performance gain in different use cases, both of which shall be considered in RAN4 in case-by-case manner.** |

## 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. Testing goals
2. Generalization agreements update
3. TR content, TR update
4. Terminology update based on R4-2320357

2nd priority

1. Latency requirements for data collection
2. Performance degradation
3. Post deployment testing
4. Ground truth
5. RAN4 recommendation regarding future work

### Sub-topic 1-1

Testing goals

The testing goals were discussed in the previous meeting (RAN4#108Bis) and several companies brought proposals based on those options. The options discussed in the previous meeting based on the moderator summary (R4-2317258) are shown below:

**Issue 1-4 from RAN4#108Bis (R4-2317258): Testing goals**

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

**Issue 1-1: Testing goals**

* Proposals
  + Option 1
  + Option 2/2a
  + Option 3
  + Option 4
* Recommended WF
  + Option 3

ZTE: we should not preclude comprehensive cases. Support option 3.

Nokia: We have issue with Option 3. We cannot test models but we can test AI/ML functionalities. Option 3, some of them are test models.

Moderator: we can have functionality test with model ID to test multiple models.

Ericsson: Option 2 can test functionality. Option 3 is OK. Rewording performance gain.

Vivo: We also think LCM is based on model ID and model switching. We still need model ID. Can we only base on functionality.

Nokia: still have model/functionality.

* Option 2a: The testing goal is to verify whether the minimum performance~~/performance gain~~ of AI/ML ~~model/~~functionality/feature can be achieved for the following ~~a static or non-static(dynamic)~~ scenario/configuration.
  + FFS how to define the static test scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)
  + FFS how to define the minimum performance target(s) (e.g., by defining AI/ML dedicated performance/core requirements associated with use cases)
  + FFS how to define the non-static specific scenarios/configurations

Agreement:

* For Testing goals, Option 1 and/or Option 2 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

### Sub-topic 1-2

*Generalization update*

Some high level agreements on generalization were reached in the last meeting and were documented in R4-2317631. These were about Issue 1-2: generalization goals and Issue 1-3: handling of generalization in tests. Some companies are proposing to make some editorial revisions to these agreements. These agreements are already capture in the draft TP proposed in R4-2318489, if changes are agree then the TP will have to be updated.

*Agreement from R4-2317631 WF:*

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

**Issue 1-2: Generalization update**

* Proposals
  + Option 1: Issue 1-2 update proposal in R4-2319824:

Apple: how can we guarantee no performance difference under various propagation conditions?

Ericsson: in principle, we have different understanding about the degradations. We do not mean that the same SNR level should be kept under various conditions.

Nokia: The propagation conditions are based on different capabilities.

Apple: eventually we need think how to test. What is the relation to static and non-static tests?

Qualcomm: degradation has complex meanings. Propagation conditions is more specific. Scenarios are not specific.

Nokia: we change it on purpose to make it concrete.

OPPO: In last meeting, we have discussed the issue. The agreement is that it is related to static condition.

Vivo: What propagation conditions mean? We propose to keep the previous agreement.

NTT DOCOMO: to OPPO, the non-static is still FFS.

Qualcomm: prefer to keep the previous agreement.

Huawei: support the previous agreement.

**Conclusion: keep the previous agreements and remove the performance gain.**

Agreement:

* Verify whether the ~~performance gain/~~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
* FFS on how to define the performance degradation
  + Option 2: Issue 1-3 update proposal in R4-2319824:
* **Modified Option 1: Signaling for Functionality-basedLCM procedures and functionality performance monitoring are considered in dedicated test cases and are excluded in tests verifying generalization and robustness. RAN4 defines multiple tests with different propagation conditions. In each of the test, TE can configure~~s~~ the same model ID when available from the UE, and therefore the same AI/ML Functionality ~~model~~ is tested under different propagation conditions to verify it’s generalizability and robustness. (environment differs in each test but not changing dynamically during the test).**
  + - **Specified UE configuration includes functionality and~~/or~~ model ID ~~if defined~~ when available from the UE**
    - **In Option 2, change the same model ID to “the same specified UE configuration, which includes functionality and~~/or~~ model ID ~~if defined~~ when available from the UE.**

Ericsson: what are the changes? Some black texts are different from the previous agreements.

Qualcomm: Even black texts include changes. We need go back to original texts and modify them. We prefer to keep the previous agreements.

Conclusion: keep the previous agreements for modified option 1 and option 2.

* + Option 3: Issue 1-2 update proposal in R4-2320610:
    - ~~FFS on details about the scenarios and/or configurations for test and the corresponding AI/ML models/functionality~~
    - The identified scenarios and/or configurations for test
      * shall include: Configuration(s) of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature/FG, for AI/ML model/functionality identification and model-ID-based/functionality-based LCM of UE-side models and/or UE-part of two-sided models;
      * shall NOT include: Additional conditions (e.g., any aspects that are assumed for the training of the model but are not a part of UE capability for the AI/ML-enabled feature/FG) as determined/identified between UE-side and NW-side

Ericsson: the proposals are premature. For the first bullet, we are OK to change to “may include”.

OPPO: Prefer to keep all the bullets. For additional condition, the channel belongs to the additional condition.

Ericsson: my concern is that “identified scenarios” does not include additional conditions.

Qualcomm: this is FFS bullet under generalization test. Shall include or shall note include prefers to the change for generalization test.

Apple: configurations mean UE configuration or something different. UE-side models and UE part of two-sided models, does network can change the configurations? Why to limit to UE side.

Ericsson: Still concern on including the second bullet. If the signaling will impact the UE performance and training, we should not include it.

* Recommended WF

To be discussed if proposed changes are agreeable

### Sub-topic 1-3

*TP Handling for TR 38.842*

Several TPs describing the RAN4 discussions and agreements so far were submitted to this meeting. Moderator proposes to take the TP submitted by the rapporteurs in R4-2318489 as baseline draft for further discussions. The goal in the meeting will be to approve the TP with any needed modifications.

**Issue 1-3: TP handling**

* Proposals
  + Option 1: Take R4-2318489 as baseline for further discussion and potential approval
  + Option 2:
  + Option 3:
* Recommended WF
  + Discuss R4-2318489 and approve

Companies are invited to provide comments and suggested changes as early as possible before/during the meeting so that there won’t be a need to spend a lot of online time updating the TP.

### Sub-topic 1-4

*Terminology update on Test encoder/decoder for TE*

A terminology update to clarify the meaning of test encoder/decoder for TE is proposed in R4-2320357.

**Issue 1-4: Terminology update**

* Proposals
  + Option 1: Agree the terminology updated proposed in R4-2320357
    - **Test encoder/decoder for TE**: AI/ML model for UE encoder/gNB decoder implemented by TE.
* Recommended WF
  + Agree Option 1

Agreement: Agree on Option 1.

### Sub-topic 1-5

*Latency requirements for data collection/inference*

Some companies are proposing to further discuss the definition of latency requirements while there are also proposals to defer this discussion to the WI phase when the use case is clear.

**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
* Recommended WF

To be discussed

ZTE: support Option3. Based on RAN2 agreement, they gave us the concrete proposal.

Qualcomm: we have already discussed it in the previous meetings. The processing time include transmission delay? T1 is just propagation delay. For delay of transmission for 1K or 1M, how can you define the formula? RAN4 does not have requirements for different data size.

Conclusion: Keep all the options and further discuss them in the WI phase.

### Sub-topic 1-6

*Peformance degradation*

Several contributions are discussing how to characterize the performance degradation and how to handle it

**Issue 1-6: Performance degradation**

* Proposals
  + Option 1: Performance degradation should be discussed case by case. And different approaches may be able to assess the degradation in one scenario/use case.
  + Option 2: Take typical validation methods which are hold-out validation, K-fold cross validation, etc
    - The dataset for validation and performance metric should be specified in the spec
  + Option 3: Take the relative performance validation
    - If the objective use case can be operated without AI, the significant degradation can be defined as "lower than without AI performance"
  + Option 4: Handle performance degradation in the LCM framework
  + Option 5: Other
* Recommended WF

To be discussed

Ericsson: we have discussed whether the degradation is allowed in the different scenarios. Hard to understand.

Moderator: this is related non-AI/ML. The degradation is for the case where UE is not trained properly.

NTT DOCOMO: all options can be taken.

Vivo: it is quite challenging. We do not know the real performance.

Apple: it is not clear to us what the context is for discussions. I do not know if we really need talk about the degradation.

### Sub-topic 1-7

*Post deployment testing*

**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
* Recommended WF

To be discussed

Qualcomm: it is unlikely to make agreements. We do not see the clear proposals for option 1 and 2.

Ericsson: from our point of view, there is need to consider the performance. It is properly Option 1.

Vivo: this is not conformance testing. It is the monitoring approach. We do not want to preclude other possible solutions.

Nokia: one intention is to have conformance testing for deployment.

### Sub-topic 1-8

*Ground truth*

RAN4 handling of ground truth was briefly discussed in the last meeting. Some contributions have further elaborated on this issue. A more in-depth discussion on this issue could be helpful

**Issue 1-8: Ground truth handling**

* Proposals
  + Option 1: Explicit definition of ground truth could be discussed further in WI use case by use case when a necessity is identified
  + Option 2: The ground truth shall be mapped to the input and compared to measurement results to guarantee the data quality.
  + Option 3: It is not possible to establish a ground truth in RAN4
  + Option 4: Other
* Recommended WF

To be discussed

Apple: what is the difference between ground truth and genie? If they are identical, we are OK with Option 1.

Nokia: this metric, we have different meaning for different metric. We prefer option 1.

Moderator: for genie, we have no clear definition. Genie is ideal results. Here we are talking about the available measurement results.

Apple: Like FR2 test, TE should know the expected results.

Moderator: TE may know the statistics.

Agreement:

* Explicit definition of ground truth could be discussed further in WI use case by use case when a necessity is identified

### Sub-topic 1-9

*RAN4 recommendation regarding future work*

R4-2319644 proposes that RAN4 makes some recommendation to plenary regarding the continuation of RAN4 work on AI/ML

**Issue 1-9: RAN4 Recommendation for future work**

* Proposals
  + Option 1: RAN4 should recommend to continue/start discussion from the beginning of any AI-PHY WI
  + Option 2: Leave such discussion to RAN plenary
  + Option 3: Other
* Recommended WF

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2318282**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318282.zip) | CATT | **Proposal 1: RAN4 not to define requirements for model delivery/update/transfer at current stage. Necessity of requirements can be studied when other WGs make progress.**  **Proposal 2: For measurement data:**   * **if data is obtained via existing measurement methods, no need to do anything.** * **if data is obtained via new measurement methods, study necessity of defining accuracy requirements case by case with the type/purpose of measurement data considered.**   **For label data:**   * **if label data are used for offline training, no need to define accuracy requirements.** * **If label data are used for monitoring, RAN4 to study the impact of label error on the model performance first.**   **Proposal 3: Discussion on reporting format for positioning sub-use case 2b and 3b should be postponed to WI phase. Study on necessity/feasibility is required after other WGs make more progress.** |
| [**R4-2318580**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318580.zip) | Apple | **Proposal 1: RAN4 shall not define requirements for model delivery/update/transfer in each use case** |
| [**R4-2318763**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318763.zip) | Nokia, Nokia Shanghai Bell | **Observation 1:** The CSI use case impacts precoding matrix part of the CSI reporting requirements.  Proposal 1: RAN4 should further study the impacts of AI/ML-enabled CSI use cases on the UE performance requirements in TS 38.101-4. A specific new target value of γ (gamma) for AI/ML-enabled CSI use cases can be envisaged.  **Observation 2:** A new relative throughput performance indicator can be introduced for AI/ML-enabled CSI use cases.  Proposal 2: RAN4 should further study if a new relative throughput performance indicator would be more suitable for AI/ML-enabled CSI use case, other than the legacy γ (gamma).  **Note: Legacy performance can be considered as baseline only for the features/use-cases that are mandatorily supported by the device.**  **Observation 3:** For CSI prediction performance monitoring, RAN1 is already discussing about network side performance monitoring option (Type 2). The same approach can be applied to test the CSI prediction accuracy as a performance KPI at RAN4.  **Observation 4:** CSI Prediction Accuracy as a KPI can be testable using the existing test interfaces with minimum change in the TE.  Proposal 3: 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.  **Observation 5:** RSRP Accuracy of RSRP prediction of Top-K beams as a KPI can be testable using the existing test interfaces, where the tolerance margin for RSRP prediction for AI/ML beam management based functionality is taken into account.  Proposal 4: RSRP accuracy of Top-1 or Top-K predicted beams should be used as performance KPI for RAN4 requirements for AI/ML based beam management.  **Observation 6:** Top-K Beam IDs prediction accuracy as a KPI can be testable using the existing test interfaces, where the predicted strongest beam ID should be within predicted Top-K beams.  Proposal 5: Top-1 or Top-K Beam IDs prediction accuracy should be used as performance KPI for RAN4 requirements for AI/ML based beam management.  **Observation 7:** Option 1, ground truth vs reported location (i.e. positioning accuracy) as a metric/KPI should be considered for UE based direct AI/ML Positioning.  **Observation 8:** Option 3 ToA, RSTD, RSRP and RSRPP as an intermediate metric/ KPI should be considered for assisted AI/ML positioning.  **Observation 9:** Option 4 LOS/ NLOS as an intermediate metric/ KPI should be considered for assisted AI/ML positioning. LOS/ NLOS metric/ KPI validation requires labelled data which can be generated based on the ground truth extracted using legacy methods.  Proposal 6: Option 4 LOS/ NLOS metric/ KPI labelled data can be generated based on the ground truth extraction using the legacy methods with the help of a calibrated device capable of precisely indicating the ratio of received LOS/NLOS signals.  Proposal 7: For BM-Case1 and BM-Case2, RAN4 should study the impact of tolerance margin in legacy L1-RSRP measurement accuracy requirements on the performance of AI/ML based BM. |
| [**R4-2318847**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318847.zip) | xiaomi | **Proposal 1: For model monitoring for CSI requirement, intermediate results, e.g. SGCS, NMSE, can be considered as metric/KPI.**  **Proposal 2: RAN4 needs to define test metric for both beam prediction accuracy and RSRP prediction accuracy.**  **Proposal 3: For L1-RSRP prediction accuracy, ideal L1-RSRP is ground-truth value without any measurement error.**  **Proposal 4: For beam prediction, both beam index based or RSRP based metric will be included in SI stage.**  **Proposal 5: All these options for Metrics/KPIs for positioning requirements/tests agreed in the previous RAN4 meeting can be included in TP. The downs election on them can be happened upon further RAN1’s conclusion during WI stage.** |
| [**R4-2319076**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319076.zip) | vivo | ***Proposal 1: Requirements and tests for model transfer/delivery is necessary to be specified and more progress from other WGs is needed.***  ***Proposal 2: RAN4 to study how to define delay requirement and accuracy requirement for model monitoring for different monitoring procedure.***  ***Proposal 3: Model monitoring requirements should be discussed per use case basis.***  ***Proposal 4: RAN4 to consider intermediate KPIs/metrics, e.g. SGCS, NMSE, for model monitoring in CSI use case.*** |
| [**R4-2319085**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319085.zip) | CMCC | ***Proposal 1: it is proposed to define requirements for model monitoring in LCM. And it is proposed to study whether to use same or different metric as that for inference per use case.***  ***Proposal 2: for CSI compression, intermediate KPI, e.g. cosine similarity, can be considered as KPI/test metrics for LCM.*** |
| [**R4-2319643**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319643.zip) | Ericsson | [Proposal 1 RAN4 to not consider *positioning accuracy: ground truth vs. reported* as one of the positioning KPIs/metrics for positioning use case.](#_Toc149912578)  [Proposal 2 RAN4 to first evaluate feasibility aspect before considering the accuracy of LoS/NLoS indication as one of the KPIs/metrics for AI/ML based positioning.](#_Toc149912579)  [Proposal 3 Accuracy of RSTD as output of AI/ML model, at least for positioning sub-use case 2a, shall be considered as one of the KPIs/metrics for positioning use case in RAN4 discussion.](#_Toc149912580)  [Proposal 4 RAN4 to not consider PRS-RSRP accuracy as one of the KPIs/metrics for positioning use case in RAN4 discussion, unless PRS-RSRP is one of the potential outputs of AI/ML model used for positioning.](#_Toc149912581)  [~~Proposal 5~~ RAN4 to discuss other positioning KPIs/metrics, if found relevant, for AI/ML based positioning during the WI phase.](#_Toc149912582)  [Proposal 6 Accuracy requirement for label data (corresponding to model output) needs to be defined if collection of training data over the air interface is agreed to be standardized.](#_Toc149912583)  [Proposal 7 Accuracy requirement for measurement data (corresponding to model input) needs to be defined.](#_Toc149912584)  [Proposal 8 RAN4 to introduce performance monitoring requirements during the WI phase.](#_Toc149912585)  [Proposal 9 RAN4 to study introducing requirements for beam pair(s) if RAN1 agreed to introduce it in WI phase.](#_Toc149912586)  [Proposal 10 Ran4 to define requirements for confidence metric calculation and reporting to make sure same baseline performance for all the UE.](#_Toc149912587)  [Proposal 11 RAN4 to study improvement to L1-RSRP measurement accuracy and the conditions under which it can be improved for model input for better model training/inference.](#_Toc149912588) |
| [**R4-2319938**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319938.zip) | OPPO | **Proposal 1: For RAN4 performance test, at least two aspects need to be considered:**   * **Model/functionality input, the focus of testing is on whether the input information of the model/functionality could be correctly obtained** * **Model/functionality output, the focus of testing is on whether the performance of a given model/functionality could be guaranteed**   **Proposal 2: For AI/ML based CSI feedback, performance requirement on CSI model/functionality input (e.g. CSI-RS measurement accuracy) should be studied.**  **Proposal 3: For CSI inference performance,** **throughput should be used to evaluate the model inference performance, and existed RAN4 test examples for “reporting of PMI” can be reused or serve as a reference. Requirement of γ and test settings can be reused or updated.**  **Proposal 4: For CSI performance monitoring, UE can monitor and estimate the performance of AI/ML based CSI model/functionality through Hypothetical BLER or intermediate KPIs, e.g. SGCS.**  **- FFS how to perform cell level AI/ML model/functionality performance monitoring**  **Proposal 5: For AI/ML based BM, performance requirement on BM model/functionality input (e.g. beam measurement accuracy) should be studied.**  **Proposal 6: For BM inference performance, RSRP accuracy of the predicted beam(s) should be used for BM model inference performance tests.**  **Proposal 7: For BM performance monitoring, UE can monitor and estimate the performance of AI/ML based CSI model/functionality through RSRP accuracy.**  **- FFS how to perform cell level AI/ML model/functionality performance monitoring**  **Proposal 8: For AI/ML based Positioning, performance requirement on Positioning model/functionality input (e.g. measurement accuracy of CIR/PDP/RSRP /RSTD) should be studied for all AI/ML positioning cases, i.e., for case 1/2a/2b/3a/3b.**  **Proposal 9: For positioning case1, positioning accuracy should be utilized as the KPI to test the model/functionality output.**  **Proposal 10: Regarding RAN4 tests on positioning accuracy, using positioning test data set(s) which including model input-related data (e.g. measurement results of PRS that matched with positions, or model inputs obtained from the measurement results) and positioning label data, prepared in advance by TE could be considered.**  **Proposal 11:** **For case2 and UE-side AI/ML assisted positioning in case1, considering the accuracy for intermediate results could be considered as the KPI to test the model/functionality output.**  **Note: most of the intermediate measurement results agreed in RAN1 are not realistic measurement results. They are from AI/ML model output(with non-linear processing). How to get the label data(the expected intermediate measurement results) and test these intermediate measurement results is also a new and a more challenge issue in RAN4. Feasibility of using these candidate metrics should be studied.**  **Proposal 12: For positioning case2b,3a,3b (cases without UE-side model), not necessary to test the Positioning model/functionality outputs.** |
| [**R4-2320185**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320185.zip) | Huawei,HiSilicon | ***Proposal 1:*** The down-selection of RAN4 identified potential test metric in AI CSI compression needs to wait RAN1/RAN2 progress on related signaling/procedure definition related to model transfer/delivery.  ***Proposal 2:*** If the model under test of the DUT is transferred over air interface signaling from the opposite site, then the throughput may not be applicable. In this case, intermediate KPIs seems to be more applicable. However, how to obtain the expected model output is still an open issue.  ***Proposal 3:*** If the model transfer/delivery of the model under test to the DUT is spec transparent, then the throughput may be applicable. However, how to eliminate the effect of the operations from the opposite side is still an open issue.  ***Proposal 4:*** RAN4 studies the potential KPIs/test metrics in AI/ML spatial-frequency CSI compression in Table 2.1.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Table 2.1 KPIs/test metrics in AI/ML Spatial-frequency CSI compression | | | | | | Test Objective | Type 1 NW Joint Training | | Type 3 Separate Training | | | gNB | UE | gNB | UE | | KPIs/Test Metrics | * Throughput * Accuracy of CSI decompression (intermediate KPIs, e.g. cosine similarity) | Accuracy of CSI compression (intermediate KPIs, e.g. cosine similarity) | * Throughput * Accuracy of CSI decompression (intermediate KPIs, e.g. cosine similarity) | * Throughput * Accuracy of CSI compression (intermediate KPIs, e.g. cosine similarity) |   ***Proposal 5:*** If throughput is the test metric for AI CSI prediction, how to ensure that the testing dataset aligns well with training dataset is still an open issue.  ***Proposal 6:*** If intermediate KPI is the test metric for AI CSI prediction, which entity provides the ideal CSI is still an open issue.   |  |  |  | | --- | --- | --- | | Table 2.2 KPIs/test metrics in AI/ML Temporal CSI prediction | | | | Test Objective | AI/ML model @ UE | | | gNB | UE | | KPIs/Test Metrics | / | * Throughput * Accuracy of CSI prediction (intermediate KPIs, e.g. cosine similarity) |   ***Proposal 7:*** RAN4 studies the potential KPIs/test metrics in AI/ML Temporal CSI prediction in Table 2.1.2. |
| [**R4-2320245**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320245.zip) | Google Inc. | **Proposal 1: Use the relative relaxation to legacy requirement to determine minimum performance for AI/ML functionality/model in identified scenarios and/or configurations and to significant degradation for AI/ML functionality/model in other scenarios and/or configurations.**  **Proposal 2: Deprioritize the generalization testing discussion on dynamically changing propagation conditions in one test case.**  **Proposal 3: Deprioritize CSI prediction accuracy as the test metric to evaluate AI/ML functionality/model inference performance for AI/ML CSI feedback enhancement use cases.**  **Proposal 4: For AI/ML functionality/model monitoring for CSI feedback enhancement use cases, it is proposed to use both throughput/relative throughput and intermediate-KPI as the performance monitoring metric.**  **Proposal 5: For test encoder/decoder at two-side model testing, prioritize option 1/option 2/option 3 and deprioritize option 4.** |

## Open issues summary

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

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. Several agreements on what to include in the TR were reached in RAN4#108Bis and documented in R4-2317631. Based on those agreements, none of the sub-topics is critical to be included in the TR. Therefore, all items are marked as 2nd priority and 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.

2nd priority:

1. Requirements for model delivery/updated/transfer
2. Measurement data
3. Labeled data
4. Metrics/KPIs for CSI performance monitoring
5. Metrics/KPIs for Beam prediction requirements/tests
6. Metrics/KPIs for positioning requirements/tests

### Sub-topic 2-1

*Model delivery/update/transfer requirements*

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

**Issue 2-1: Model delivery/update/transfer requirements**

* Proposals
  + Option 1: Do not discuss requirements further at current stage, further discuss/study based on progress in other groups
  + Option 2: There is no need for such requirements
  + Option 3: Others
* Recommended WF

Option 1

### Sub-topic 2-2

*Accuracy requirements for measurement data*

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

**Issue 2-2: Accuracy requirements for measurement data**

* Proposals
  + Option 1: Handling of measurement data requirements:
    - if data is obtained via existing measurement methods, no need to do anything.
    - if data is obtained via new measurement methods, study necessity of defining accuracy requirements case by case with the type/purpose of measurement data considered.
  + Option 2: Acccuracy requirements will be needed, further discuss in a possible WI
  + Option 3: Other
* Recommended WF
  + To be discussed

### Sub-topic 2-3

*Accuracy requirements for labeled data*

Some papers are discussing the possibility/necessity to define requirements for labeled data

**Issue 2-3: Accuracy requirements for labelled data**

* Proposals
  + Option 1: Handling of labelled data requirements:
    - if label data are used for offline training, no need to define accuracy requirements.
    - If label data are used for monitoring, RAN4 to study the impact of label error on the model performance first.
  + Option 2: Acccuracy requirements will be needed, further discuss in a possible WI
  + Option 3: Others
* Recommended WF
  + To be discussed

### Sub-topic 2-4

*Metrics/KPIs for CSI performance monitoring*

In the previous meeting there was an agreement in R4-2317631 that throughput will be the baseline while other KPIs are still to be considered and will be documented in the TR.

**Issue 2-4: Metrics/KPIs for CSI monitoring**

Proposals

* + Option 1: SGCS, NMSE can be considered for monitoring
  + Option 2: Hypothetical BLER
  + Option 3: Other
* Recommended WF
  + To be discussed

### Sub-topic 2-5

*Metrics/KPIs for Beam prediction requirements/tests*

Agreement to capture all the options in the TR. Some companies are proposing to further down select

**Issue 2-5: Metrics/KPIs for Beam prediction requirements/tests**

* Proposals
  + Option 1: Use the following metrics as performance KPI for RAN4 requirements for AI/ML based beam management:
    - RSRP accuracy of Top-1 or Top-K predicted beams
    - Top-1 or Top-K Beam IDs prediction accuracy
  + Option 2: Do not do any further downselection, discuss further in the WI phase
* Recommended WF
  + To be discussed

### Sub-topic 2-6

*Metrics/KPIs for positioning requirements/tests*

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

**Issue 2-6: Metrics/KPIs for positioning requirements/tests**

* Proposals
  + Option 1: Consider LOS/ NLOS metric/ KPI,
    - labelled data can be generated based on the ground truth extraction using the legacy methods with the help of a calibrated device capable of precisely indicating the ratio of received LOS/NLOS signals
  + Option 2: Do not do any further downselection in the SI stage
  + Option 3: Do not consider positioning accuracy: ground truth vs. reported as one of the positioning KPIs/metrics for positioning use case
  + Option 4: Consider accuracy of RSTD as output of AI/ML model, at least for positioning sub-use case 2a, as one of the KPIs/metrics for positioning use case in RAN4 discussion.
  + Option 5:Do not consider PRS-RSRP accuracy as one of the KPIs/metrics for positioning use case unless PRS-RSRP is one of the potential outputs of AI/ML model used for positioning.
  + Option 6: For positioning case1, positioning accuracy should be utilized as the KPI to test the model/functionality output.
  + Option 7: For case2 and UE-side AI/ML assisted positioning in case1, considering the accuracy for intermediate results could be considered as the KPI to test the model/functionality output.
  + Option 8: For positioning case2b,3a,3b (cases without UE-side model), not necessary to test the Positioning model/functionality outputs.
* Recommended WF
  + To be discussed

Options are not mutually exclusive, multiple options can be picked to continue studying

# Topic #3: Interoperability and testability aspect

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2318251**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318251.zip) | CAICT | **Proposal 1: If option 4 is adopted, test interoperability should be ensured with partially specified parameters.**  **Proposal 2: If option 4 is adopted, a partially standardized dataset for training of the test decoder could be considered. The content needs to be specified can be discussed in WI.**  **Observation 1: There is no need to mandate decoder sharing in the test or specification.**  **Proposal 3: Our views on comparison table of test decoder for 2-sided model are listed in the table 1.**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **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) | Up to decoder implementer (TE vendor). Alignment with UE/gNB vendors 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) | No direct correspondence | No direct correspondence | FFS | No direct correspondence | | Test decoder verification procedure at TE and/or DUT | TE needs to verify the decoder runs properly. | DUT and TE need to verify the decoder runs properly | Not needed | Needed for DUT | | Feasibility of test decoder verification procedure | Yes | Not clear yet | Not needed | Yes | | **Pros/Cons analysis** | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | May not reflect the actual decoder implemented by gNB vendor | Yes.  The alignment between TE and DUT needs FFS. | May not reflect the actual decoder | May not reflect the actual decoder. | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Medium. TE needs to cope with multiple decoders from multiple UE vendors | Medium. TE need to implement multiple decoders from different gNB vendors | Low | High. Model training should be supported by TE vendor. | | Specification Effort (e.g. test decoder) | Low | Low | High | High | | Confidentiality/IP issues | Model exposure is required from UE to TE | Model exposure is required from gNB to TE(or DUT?) | No | Not necessarily required. This is up to the collaboration between TE vendor and DUT/infra vendor. | | Applicability to different scenarios/conditions/ configurations | Yes | Yes | Conditional Yes. | Yes. | | Complexity of actual testing procedure for the ecosystem | Medium. Depend on the verification procedure between DUT and TE. | High. Depend on the verification procedure between DUT and TE/gNB. | Low | Medium, Depend on the verification procedure between DUT and TE. |   **Proposal 4: Updated reference block diagram of 1-sided model in Figure 2 could be considered.**  **DUT**  **Model/Functionality monitoring**  **Model inference**  **Model/Functionality select/switch/update/activate/deactivate**  **Data collection**  **(Inference/monitoring)**  **input**  **output**  **Data for monitoring**  **Control**    **TE**  **Test controller**  **Scenario generator (channel model, conditions, etc.)**  **Conductive or OTA link**  **Test signaling generation**  **Performance verification**  **Model/Functionality LCM control**  **Data collection**  Figure 2 Updated reference block diagram of 1-sided model  **Proposal 5: Updated reference block diagram of 2-sided model in Figure 3 could be considered.**  **DUT**  **Model/Functionality monitoring**  **Model inference**  **Model/Functionality select/switch/update/activate/deactivate**  **Data collection**  **(Inference/monitoring)**  **input**  **Data for monitoring**  **Control**  **TE**  **Test controller**  **Scenario generator (channel model, conditions, etc.)**  **Conductive or OTA link**  **Test signaling generation**  **Performance verification**  **Model/Functionality LCM control**  **Model inference**  **output**  **output**  Figure 3 Updated reference block diagram of 2-sided model |
| [**R4-2318283**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318283.zip) | CATT | **Proposal 1: The following parameters may be specified for test decoder by RAN4:**   * **Model structure** * **Activation function** * **Maximum FLOPs allowed for the test decoder** * **Maximum number/size of parameters** * **Others**   **Proposal 2: It is considered similar among possible implementations when the chosen metrics/KPIs are comparable. The specific values of difference between possible implementation can be FFS in WI phase after metrics/KPIs are stable.**  **Proposal 3: RAN4 not to define a standardized data set for the specified test decoder.**  **Proposal 4: In general, the decoder is not shared with DUT vendors and infra vendors due to confidentiality issue unless the decoder is trained with common data set or the DUT/TE vendors are willing to share it.** |
| [**R4-2318479**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318479.zip) | NTT DOCOMO, INC. | **Proposal 1: The data set used for the decoder training should be standardized to bound the output variation within certain range and ensure the repeatability.**  **Proposal 2: If data set is standardized and other vendors can develop a decoder which can deliver similar performance, DUT vendors and infra vendors can prepare the reference decoder and performance is similar, thus the decoder developed by TE vendor is not needed to be shared with them.**  **Proposal 3: For some issues, we provide our opinion in the table below (yellow highlighted):**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **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 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) | It should be specified to bound the output variation within certain range and ensure the repeatability | | 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) | No collaboration is defined because DUT vendor develops both encoder and decoder, or this can be included in all the Types | Type 1 or 3. It depends on source of decoder training data. | No collaboration is defined because decoder model is fully specified, or this can be included in all the Types | Type 3 according to the assumption that training data set is standardized and decoder model is not needed to be shared. | | Test decoder performance verification procedure at TE and/or DUT | TE side verification is needed. Since the decoder is developed by DUT, it should be verified at TE | TE side verification is needed. Since the decoder is developed by DUT, it should be verified at TE | Not needed because decoder model is fully specified | Not needed because TE develops the decoder | | Feasibility of test decoder verification procedure |  |  |  |  | | Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) |  |  |  |  | |  | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | Probably low. It depends on the training data | Probably high. It depends on the training data | Low. It is difficult to reflect the real deployment to standardized model | Probably low. . It depends on the training data | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) |  |  |  |  | | Specification Effort (e.g. test decoder) |  |  |  |  | | Confidentiality/ IP issues |  |  |  |  | | Applicability to different scenarios/conditions/ configurations | It depends on generalization test procedure | It depends on generalization test procedure | It depends on generalization test procedure | It depends on generalization test procedure | | Complexity of actual testing procedure for the ecosystem |  |  |  |  | | Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes, DUT vendor can update decoder model | Yes, decoder implementer can update decoder model | No, it needs discussions how to specify | Yes, TE vendor can update decoder model | | Relationship with reference decoder/encoder for defining requirement | Reference encoder/decoder is not needed to be considered according to the agreements in ad-hoc session in #108 [2] | Reference encoder/decoder is not needed to be considered according to the agreements in ad-hoc session in #108 [2] | Reference encoder/decoder is not needed to be considered according to the agreements in ad-hoc session in #108 [2] | Reference encoder/decoder is not needed to be considered according to the agreements in ad-hoc session in #108 [2] | | Whether model transfer/delivery is needed during the test procedure | Not needed | Not needed | Not needed | Not needed | |
| [**R4-2318581**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318581.zip) | Apple | **Observation 1: With standardized data set:**  **Pros: Reduce variability in the performance across different TE implementations.**  **Cons: Complexity to collect data, reduces implementation flexibility, mismatches with real deployment,**  **Observation 2: Data set not standardized:**  **Pros: Accommodate the different in implementation flexibility, reduce mismatches with real deployment.**  **Cons: Large performance variations, performance will depend on specific dataset from different NW vendors**.  **Observation 3: UE vendor or NW vendor can potentially provide the training data set with Type1 or Type3 collaboration**  **Proposal 1: RAN4 further discuss whether the training data set should be standardized based on the above pros and cons**  **Proposal 2: RAN4 further clarify:**   * **Why does the test decoder need to be shared with DUT or infra vendor?** * **Is the proposed decoder to be shared with DUT before or after training?**   **Proposal 3: RAN4 further clarify in option 4:**   * **Is the decoder and encoder (DUT) being trained at TE before testing?**   Table 1: Options analysis for test decoder(previous agreements marked in yellow)   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Option 1: DUT provides decoder** | **Option 2: Decoder not from DUT and Spec** | **Option 3: Full decoder specification in standard** | **Option 4: partially**  **specified decoder** | | Clarification of options | | | | | | Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | RAN4 specification | TE vendor, decoder  implementation based on  RAN4 specifications | | Source of decoder training data | Up to DUT 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) | For type 1 UE side and type 3 UE first the training data are provided by UE vendor. For type 1 NW side and type 3 NW first the training data are provided by NW vendor. | | 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) *\* Note: RAN4 specification of training collaboration procedure before the test is not needed and simulation assumption on training/collaboration type can be discussed separately in the WI stage.* | Type1 with UE side training  Type3 with UE first training. UE provides the reference decoder as the test decoder to TE | Type1 with NW side training  Type3 with NW first training  Type2 depending on offline alignment between DUT vendor and decoder implementation. | All three types are possible if RAN4 is successful in specifying a decoder. | Any type of collaboration is possible depending on the alignment of the TE and DUT | | Test decoder verification procedure at TE and/or DUT | Up to the UE to guarantee the decoder works with the encoder. Tested at DUT | The reference encoder used by NW should be made available to verify the decoder provided by NW. | Should be guaranteed when specified, no need to test | Needed (it needs to be proved that the decoder works such that any test failure is attributed to DUT only) | | Feasibility of test decoder verification procedure | Verification feasible at DUT | Feasible at TE with reference encoder from NW | Feasible during specification effort | Feasible at TE with reference encoder from NW if the training data is provided by NW vendor  Feasible at TE with reference encoder from UE if the training data is provided by UE vendor | | Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) |  |  |  |  | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | Low  It could potentially be the same model used in testing and deployment with Type 1 UE side training  But Model mismatches may happen in the field if network vendor would not use the test decoder (that was provided by the UE vendor) and was used during testing.  There are no model mismatches during the test. UE perfectly knows the decoder since it was designed by UE. This could result in UE passing the tests easily but fail in the field. (mismatch with a different decoder not designed by UE) | Low  It could potentially be the same model used in testing and deployment with Type 1 NW side training  The encoder(s) of UE vendor will need to match all the reference decoders provided by different network vendor. This could cause mismatches during the test as well as in the field. | Low  It could limit implementation of encoder model to work with the test decoder, or potentially lead to mismatch in model used for testing and in deployment  An encoder that passes the test with the test decoder may not work for the decoder in the field.  There is less flexibility in implementation with fully specified reference decoder. There could be a mismatch with the decoder in the field. UE also loses its flexibility for implementing the encoder. UE can implement an additional encoder for the purpose of RAN4 test only. | Low  With no knowledge of the full decoder, there might be a mismatch in encoder-decoder pair, as test decoder is only partially specified. It could limit implementation of encoder model to work with the test decoder, or potentially lead to mismatch in model | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | High  UE vendor provides trained model to TE.  TE has to support all the test decoders provided by different UE vendors | High  NW vendor provides trained model to TE  TE needs to first verify the test decoder before it will be used for DUT testing  TE has to support all the test decoders provided by different network vendors | High  Assume no training needed, fully specified. Complex to ensure inter-operability with different UE vendor encoder models | High  Complex to ensure inter-operability with different UE vendor encoder models  Different performance may be achieved by different decoders implemented by TE vendors. | | Specification Effort (e.g. test decoder) | Need not be specified | Need not be specified | High  It would take a lot of time and effort to reach consensus on a test decoder model to be specified by RAN4 | High  It would take a lot of time and effort to reach consensus on a test decoder model to be specified by RAN4 | | Confidentiality/ IP issues | Disclose to TE | Disclose to TE | Decoder already specified, and disclosed | Decoder already specified, and disclosed | | Applicability to different scenarios/conditions/ configurations | The model should be applicable for the scenarios/ configurations tested for in RAN4 | The model should be applicable for the scenarios/ configurations tested for in RAN4 | The model should be applicable for the scenarios/ configurations tested for in RAN4 | The model should be applicable for the scenarios/ configurations tested for in RAN4 | | Complexity of actual testing procedure for the ecosystem | Complex | Complex | Complex | Complex | | 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 for defining requirement | [Alt 1: same as reference decoder  May not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.  Alt 2: different from reference decoder  UE may not pass the tests due to different test decoders are used for defining requirements and tests.]  More clarifications is needed | [Alt 1: same as reference decoder  May not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.  Alt 2: different from reference decoder  UE may not pass the tests due to different test decoders are used for defining requirements and tests.] | [Alt 1: same as reference decoder  Possible to define requirements and be able to calibrate results from companies.  Alt 2: different from reference decoder  There is no reason to specify test decoder different from that is used for defining requirements. ] | [Alt 1: same as reference decoder  There is good chance that the results among companies can be calibrated as the performance of the model could largely be decided by the specified part.  Possible to define requirements  Alt 2: different from reference decoder  There is no reason to specify different test decoder than that is used for defining requirements.] | | Whether model transfer/delivery is needed during the test procedure | Yes | Yes | No | No |   **Observation 4: Building a test dataset generated from the field poses many challenges on the effort and quality of data.**  **Proposal 4: Exclude CDL channel model from testing data**  Channel models for testing  **Proposal 5: Study the feasibility of current models employed in RAN4 requirements for testing AI/ML. Based on this feasibility we could explore other options if necessary**  **Proposal 6: Investigate the feasibility of defining tests that access the AI/ML performance under a wide range of SNR conditions while also keeping the testing burden low** |
| [**R4-2318764**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318764.zip) | Nokia, Nokia Shanghai Bell | **Observation 1:** The test decoder design will become much simpler if the stake holders can share their training data.  **Proposal 1: RAN4 should further discuss the collaborative approach to test decoder design where the training data is shared by all the stake holders.**  **Observation 2:** Following parameters from Table 1 should be considered to ensure repeatability of the Tests - Encoder Input type, Decoder Input size per Rank, Decoder Output size/compression ratio per Rank, Quantization level, Type of Quantization, Supported Ranks.   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Option 1** | **Option 2** | **Option 3** | **Option 4** | | **Performance Parameters** | * Cosine similarity threshold * γ Threshold value *(Note 1)* | * Cosine similarity threshold * γ Threshold value | * Cosine similarity threshold * γ Threshold value | * Cosine similarity threshold * γ Threshold value | | **Implementation Parameters** | * Encoder Input type *(Note 2)* * Decoder Input size per Rank * Decoder Output size/compression ratio per Rank * Quantization level * Type of Quantization * Minimum Supported Ranks | * Encoder Input type * Decoder Input size per Rank * Decoder Output size/compression ratio per Rank * Quantization level * Type of Quantization * Minimum Supported Ranks | * Encoder Input type * Decoder Output size/compression ratio per Rank * Quantization level * Type of Quantization * Model size *(Note 3)* * Model architecture details *(Note 4)* * Training Dataset * Model training types * Supported Ranks | * Encoder Input type * Decoder Input size per Rank * Decoder Output size/compression ratio per Rank * Quantization level * Type of Quantization * Maximum Model size * Training Data-set size * Training Data-set details * Model training types * Minimum Supported Ranks |   *Table 1: Parameters that can be specified for a test decoder at RAN4*  **Proposal 2: RAN4 should further discuss the test decoder parameters from Table 1 for specification at RAN4.**  **Observation 3:** Based on the discussions and agreements in RAN1 and RAN2, it is more natural for RAN4 to first study the ML-enabled Functionality/Feature related inter-operability aspects, and only later address the ML model related aspects if needed.   |  |  |  |  | | --- | --- | --- | --- | |  | ML Training | Functionality based LCM (configuration/(de)activation/ switching/fallback) | Functionality/Feature performance | | NW-UE Collaboration  Level-y | N/A in Release 18 (training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | Interoperability to be guaranteed by:  - Testing of Functionality-based LCM procedures | Interoperability to be guaranteed by: - Pre-deployment Functionality performance monitoring and validation. - Post-deployment Functionality performance monitoring and validation. |   *Table 3: RAN4 interoperability analysis with focus on ML-enabled Functionality/Feature related aspects for Release 18 use cases.*  **Proposal 3: RAN4 to adopt and capture in the TR the interoperability analysis Table 3 with focus on ML-enabled Functionality/Feature related aspects for Release 18 use cases.**  **Observation 4:** A special channel model is needed instead of just TDL models that are currently used. More advanced models can be considered for the test such as CDL. Another aspect is time evolution of the channel model.  **Proposal 4: RAN4 to study whether TDL models are sufficient for the performance evaluation of AI/ML Enabled CSI feedback use-cases.** |
| [**R4-2318848**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318848.zip) | xiaomi | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | 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) | Alignment of training data between UE and gNB 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 is 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 | Needed | Needed | No. | Needed. | | 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. | | Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | Depending on the generation of the model | Depending on the generation of the model | Depending on the generation of the model | Depending on the generation of the model | | 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) | High, TE needs to implement multiple decoders from different vendors | High, TE needs to implement multiple decoders from different vendors | Low | Depends on what’s partially specified for the decoder. Offline alignment may still be needed. | | Specification Effort (e.g. test decoder) | Low, if the decoder is up to implementation, there is no spec impact | Low, if the decoder is up to implementation, there is no spec impact | High, may results in long discussion. Too many issues need to be converged. For example:   * FFS: Decoder size * FFS: Detail Decoder structure * FFS: How many decoders, how many typical scenarios | Depends on what’s partially specified for the decoder | | Confidentiality/IP issues | Need to be considered | Need to be considered | No. | Depends on what’s partially specified for the decoder | | 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 | Depend on the verification procedure between UE and TE. | Depend on the verification procedure between UE and TE/gNB. | Low | Depend on the verification procedure between UE and TE/gNB. | | 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 more simple 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. | |
| [**R4-2318936**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2318936.zip) | Qualcomm, Inc. | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **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 | | | | | | 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) |  | | 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) | No need to discuss | | | | | Test decoder verification procedure at TE and/or DUT | [To verify the test repeatability (variation among TE vendor implementations are bounded) of the test decoder] | [To verify the test repeatability (variation among TE vendor implementations are bounded) of the test decoder] | No need | To verify the test repeatability (variation among TE vendor implementations are bounded) of the test decoder | | Feasibility of test decoder verification procedure |  |  |  |  | | Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | No need to discuss if test repeatability can be ensured as agreed in the previous meeting | | | | | 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 | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Highest requirements for TE since TE needs to support multiple test decoders own by other entities for each test configuration. | Higher requirements for TE since TE needs to support one or multiple test decoders own by other entities for each test configuration. | Lower requirements for TE since TE only needs to support one standardized test decoder for each test configuration | Lowest requirements for TE since TE only needs to support one test decoder designed by itself for each test configuration | | Specification Effort (e.g. test decoder) | See proposal 6,7 | | | | | Confidentiality/ IP issues | Yes | Yes | No | No | | Applicability to different scenarios/conditions/ configurations | No difference across scenarios | | | | | Complexity of actual testing procedure for the ecosystem | Comparing the options with legacy test procedure, evaluate additional works/steps on other entities after RAN4 finish specification until finish certification   * + - * RAN5 specification       * TE vendor implementation considerations       * DUT vendor test execution       * Network (other vendor) test execution | | | | | 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 | See proposal 9 | | | |   **Proposal 6: “*Specification Effort*” for each option are analyzed based on the necessity of the following different types of specification efforts:**   * **Effort to specify high level parameters** * **Effort for designing specifications to ensure test repeatability** * **Effort for designing specifications to ensure ability to implement decoders with similar performance by other vendors** * **Effort to specify decoder performance verification procedure** * **Effort to fully specify the decoder**   To better analyze the specification efforts in each category, we provide examples/options for how to derive the specification in some of the categories.  **Proposal 7: We provide examples/potential options for derivation of the specification in some of the above categories for test decoder options:**   * **Effort for designing specifications to ensure test repeatability and ability to implement decoders with similar performance by other vendors**   + **E.g. (or option A): Capture dataset with (nominal encoder input, latent message as encoder output) in the specification for TE vendors to train the decoder, to ensure test repeatability and ability for other vendors to train the decoder with similar performance.**      - **FFS the definition of nominal encoder input, e.g., eigen vectors**     - **FFS how to derive the dataset** * **Effort to fully specify the decoder**   + **E.g. (or option A): Decide one reference encoder structure, and each company perform joint training of its encoder/decoder pair based on the reference structures and the agreed high-level parameters. RAN4 then pick an encoder/decoder pair based on the agreed evaluation and selection criterion**      - **FFS evaluation and selection criterion**  |  |  |  |  |  | | --- | --- | --- | --- | --- | | Specification effort category | Option 1: DUT provided decoder | Option 2: NW provided decoder | Option 3: Fully specified decoder | Option 4: TE provided decoder | | High level parameters | Yes | Yes | Yes | Yes | | Ensure test repeatability | (To be discussed) | (To be discussed) | No | Yes,  E.g. (or option A): Capture dataset with (nominal encoder input, latent message as encoder output) in the specification for TE vendors to train the decoder, to ensure test repeatability and ability for other vendors to train the decoder with similar performance.  FFS the definition of nominal encoder input, e.g., eigen vectors  FFS how to derive the dataset | | Ensure ability to implement decoders with similar performance by other vendors | (To be discussed) | (To be discussed) | No | | Decoder performance verification procedure | (To be discussed) | (To be discussed) | No | (To be discussed) | | Full specification of the test decoder | No | No | Yes,  E.g. (or option A): Decide one reference encoder structure, and each company perform joint training of its encoder/decoder pair based on the reference structures and the agreed high-level parameters. RAN4 then pick an encoder/decoder pair based on the agreed evaluation and selection criterion  FFS evaluation and selection criterion | No |   **Proposal 9: 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 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**   **Proposal 10: Option 1 and 2 are feasible if the loaded file representing test decoder isn’t subject to IP constraint and TE is able to load them correctly.**  **Proposal 11: Option 3 is feasible since there is a proper decoder derivation procedure as described in proposal 7 for pro & cons analysis, without IP issue and low TE requirements.**  **Proposal 12: Option 4 is feasible since it doesn’t have IP issue and with lowest TE requirement. In additional, the two additional properties, test repeatability and implement ability of other vendors, can be achieved at least by the example described in proposal 7 for pro & cons analysis.** |
| [**R4-2319077**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319077.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.***  ***Proposal 4: High-level parameters to be specified for decoder for all the options would at least include model structure (back-bone, parameters, e.g., number of layers, model (parameter) size, etc.) and processing complexity (FLOPS).***  There were also extensive discussions in the last meeting on the analysis and some agreements were reached which is highlighted in Table 1. However, there are still lots of conclusions to be made on the analysis. Updated summary is provided in Table 3.  **Table 3 Summary of 4 options for testing of 2-sided model**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **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 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) | Alt 1: RAN4 specifications  Alt 2: Up to decoder implementer (TE vendor)   * FFS whether alignment with UE/gNB vendors is 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) | 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 verification procedure at TE and/or DUT | Needed | Needed | Needed | Needed | | Feasibility of test decoder verification procedure | A reference encoder is used and it passes the test(s) with the test decoder | A reference encoder is used and it passes the test(s) with the test decoder | A reference encoder is used and it passes the test(s) with the test decoder | A reference encoder is used and it passes the test(s) with the test decoder | | Pros/Cons/Feasibility analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | 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/High  Could reflect the performance in the field since network vendors may use same or similar decoder in the field as the test decoder. | 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/High  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, interoperability) | High  TE has to support multiple test decoders provided by different UE vendors.  The test decoder should be provided in open format. Otherwise, TE will need to support a wide range of model structures and interfaces, and one to many offline co-engineering is needed.  No additional training required by TE vendor. | High  TE has to support multiple test decoders provided by different infra vendors.  The test decoder should be provided in open format. Otherwise, TE will need to support a wide range of model structures and interfaces, and one to many offline co-engineering is needed.  No additional training required by TE vendor. | Low  Low complexity for TE vendors to implement the fully specified test decoder. | Medium  Low complexity for TE vendors to implement the partially specified test decoder.  Training is needed to finalize the test decoder by TE vendors. | | Specification Effort (e.g. test decoder) | Low  specify some high-level parameters for the decoder (e.g., parameters related to processing complexity, model structure, etc) | Low  specify some high-level parameters for the decoder (e.g., parameters related to processing complexity, model structure, etc) | High  A fully specified decoder that can be directly used (e.g., model structure, model parameters, model format etc)  Training dataset to be used. | Medium to high  A partially specified decoder to be further trained (e.g., parameters related to processing complexity, model structure, etc)  How to ensure test repeatability.  Maybe training dataset to be used. | | Confidentiality/ IP issues | 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 | None  Fully specified decoder is captured in the specifications publicly. | None  Partially specified decoder is captured in the specifications publicly. TE vendors will train and finalize test decoder with the partially specified decoder. | | Applicability to different scenarios/conditions/ configurations | Applicable  Depending on how generalization test is defined | Applicable  Depending on how generalization test is defined | Applicable  Depending on how generalization test is defined | Applicable  Depending on how generalization test is defined | | Complexity of actual testing procedure 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. | Low/Medium  TE only needs to train and implement partially specified test decoder.  There could be different performance from TE vendors and DUT is supposed be tested against equivalent TE vendor implementation of the test decoder. | | 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 for defining 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, or different reference decoder based on option 3 for defining requirements. | | Whether model transfer/delivery is needed during the test procedure | No | No | No | No |   ***Proposal 5:*** ***Take into consideration the summary of 4 options for testing of 2-sided model in Table 3.***  ***Proposal 6: Reference block diagrams in Fig 9 and Fig 10 for one-sided model and 2-sided model, and functional block description in Table 4 are used for test framework for AI/ML.***  Table 4 Description of reference functional blocks   |  |  | | --- | --- | | Functional block | Description | | DUT | Device under test. It can be UE or gNB. | | Test system | A system to test AI/ML functionality/performance. It may be test equipment or gNB in practical NW. | | Test setup | Setup test environment based on design of test cases | | Data generator | This function is to generate test dataset for the ongoing test. | | AI/ML model control | In tests for verifying model inference performance, AI/ML model control may be used for model selection, and model activation if necessary.  In tests for LCM procedure, AI/ML control may be used for model selection, switch, activation, deactivation, transfer, delivery, update or model monitoring | | Test model | Test decoder/encoder for UE and gNB, respectively for 2-sided model. | | Performance requirements verification | This function is to verify if the performance requirements for a test can be met in the ongoing test. | | LCM requirements verification | This function is to verify if the LCM related requirements for a test can be met in the ongoing test. |     Fig. 9. Reference block diagram for one-sided AI/ML model    Fig 10. Reference block diagram for 2-sided AI/ML model |
| [**R4-2319086**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319086.zip) | CMCC | ***Proposal 1: for generalization test, it is proposed to consider changing environment during the test.***  ***Proposal 2: for 2-sided model testing options, the consideration are provided as following:***   |  |  |  |  | | --- | --- | --- | --- | |  | **Option 1: DUT provides decoder** | **Option 2: Decoder not from DUT and Spec** | **Option 3: Full decoder specification in standard** | | Source of the test decoder | DUT vendor | Decoder vendor (infra vendor in case of testing UEs) | 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) | | 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 | | Supported training collaboration type (source of training data should be consistent with the collaboration type) | Type 1 | Type 3 | N/A | | Test decoder verification procedure at TE and/or DUT | Needed at TE, and how to verify is FFS | Needed at TE, and how to verify is FFS | Not needed | | Feasibility of test decoder verification procedure | Feasible | Feasible | N/A | | Pros/Cons analysis | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | Low, since the test decoder may be mismatch with the decoders deployed in the field, and UE may easily pass the test since UE has full knowledge of the decoder | High | Low, since the test decoder may be mismatch with the decoders deployed in the field, and UE UE may easily pass the test since UE could train the model based on the specified decoder | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | High, since TE may need to implement multiple docoders from different vendors | High, since TE may need to implement multiple docoders from different vendors | Low | | Specification Effort (e.g. test decoder) | Low, if the decoder is up to implementation, there is no spec impact | Low, if the decoder is up to implementation, there is no spec impact | High, since the decoder(s) are fully specified and captured in RAN4 spec, may results in long discussion | | Confidentiality/IP issues | Need to be condiered | Need to be condiered | Since the decoder(s) are fully specified and captured in RAN4 spec, no IP issues | | Applicability to different scenarios/conditions/ configurations | This is pending on how to design the test to guarantee the generalization | This is pending on how to design the test to guarantee the generalization | This is pending on how to design the test to guarantee the generalization | |
| [**R4-2319642**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319642.zip) | Ericsson | ***Observation 1 In the RAN4 context, generalization should refer to the ability of the AI functionality to provide a minimum expected performance in all expected scenarios, regardless of training.***  ***Observation 2 It is necessary to identify the range of scenarios in which minimum performance is to be expected, a subset of scenarios within which to define requirements.***  ***Observation 3 The requirements may differ in different scenarios due to different conditions. However, differing performance due to differing conditions is different to degradation (which is performing worse than the optimum for a condition).***  ***Observation 4 It is necessary to determine how the AI functionality performance varies with changing scenarios.***  ***Observation 5 In some cases, it may be possible to define sub-sets of scenarios/requirements depending on UE capability or signaling outside capability framework.***  ***Observation 6 Although RAN based LCM is not considered in the performance requirements, it may still be that the UE switches model depending on scenario.***  ***Observation 7 It is not clear whether issue 1-6 and the conclusion is referring to observation of specific models or performance monitoring of the AI functionality.***  ***Observation 8 Measurement of performance of a UE against RAN4 metrics in the field would be by means of RAN1/RAN2 procedures, not test equipment.***  ***Observation 9 It is not envisaged that compliance testing would be performed in the field. Development of standardized metrics for comparing performance may be of interest, but minimum requirements may not be.***  ***Observation 10 If measurement reports are provided to the network for monitoring the performance of an AI functionality, RAN4 should consider how accurately the measurement reports would really relate to ground truth.***  ***Observation 11 Even if the reported information does not correspond with exact accuracy to ground truth, it may still be useful for a coarse metric on model performance.***  ***Observation 12 If performance monitoring consists of standardized information being periodically sent to the network to assess an assumed metric, the extent of RAN4 involvement should be discussed. It may be for the metric itself, or just the accuracy requirements on the reported information.***  ***Observation 13 If the UE reports performance or reliability information, RAN4 could set requirements on the minimum accuracy for the performance/reliability report.***  ***Observation 14 Testing of a requirement on the accuracy of reliability / performance reporting could be done in a test-house as part of compliance testing, since TE could generate ground truth and assess the real reliability/performance.***  ***Observation 15 It may not be straightforward for a test specification to force different levels of reliability/performance for testing a metric.***  Based on the discussion in the previous sections we propose the following:  [Proposal 1 To define AI requirements, the following steps are needed:](#_Toc149903848)  [ Identify the range(s) of scenarios over which the AI is expected to achieve a minimum performance.](#_Toc149903849)  [ This may involve some consideration of the range of scenarios over which non-AI is expected to perform.](#_Toc149903850)  [ Identify which scenarios the AI should show optimum performance and whether there are any scenarios in which performance may be degraded compared to the optimum for that scenario.](#_Toc149903851)  [ Note that optimum performance may differ depending on scenario. Degraded performance is performance that is less than the optimum for that scenario.](#_Toc149903852)  [ Determine how the AI performance changes depending on the scenario, whether the performance change is smooth or not.](#_Toc149903853)  [ Determine a set of conditions in which to define RAN4 requirements.](#_Toc149903854)  [ Simulate and determine requirements for each of the conditions.](#_Toc149903855)  [ Determine whether the requirements should be grouped into sub-sets with capability signaling (or other signaling).](#_Toc149903856)  [Proposal 2 If the UE switches model outside of RAN LCM, it should not impact performance.](#_Toc149903857)  [Proposal 3 RAN4 should discuss whether option 1 can ensure that physical UE model switching is transparent.](#_Toc149903858)  [Proposal 4 Assume that performance monitoring requirements in the field would apply for functionality monitoring.](#_Toc149903859)  [Proposal 5 RAN4 discuss further the feasibility of a requirement and test for UE reliability/performance reporting for monitoring of UE sided models.](#_Toc149903860) |
| [**R4-2319939**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2319939.zip) | OPPO | **Proposal 1: Regarding the testability of two-sided model, should introduce test encoder(s) to collaborate with the NW decoder under test.**  **Proposal 2: Regarding the testability of two-sided model, should introduce test decoder(s) to collaborate with the UE encoder under test.**  **Observation 1: Pros and cons for different options on test decoder are shown in table 1.**  **Table 1 pros and cons of different options for test decoder**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Option 1 | Option 2 | Option 3 | Option 4 | | Clarification of options | | | | | | Source of the test decoder | 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) | RAN4 specifications  Alignment with UE/gNB vendors is required to avoid mismatch issue between the test decoder and the encoder under test | | 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) | *In general, the way a model is trained can be decoupled from how the model is tested.*  *However, for some training types, there are specific testing methods that can more closely match the deployment scenario.*  *For example, for option1 and option2,*  *(1) with type1 UE sided 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 sided 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 provides 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.* | | | | | Options:   1. Training up to DUT vendor. The way a model is trained can be decoupled from how the model is tested. 2. Only for Type1 UE sided training, and Type3 UE first training | Options:   1. Training up to DUT vendor.The way a model is trained can be decoupled from how the model is tested. 2. Only for Type1 NW sided training, and Type3 NW first training | Encoder training up to DUT vendor. The way a model is trained can be decoupled from how the model is tested. | Training up to DUT vendor. The way a model is trained can be decoupled from how the model is tested. | | Test decoder verification procedure at TE and/or DUT | Needed | Needed | Needed | Needed | | Feasibility of test decoder verification procedure | FFS | FFS | FFS | FFS | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | Low, 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. The limited number of test models can not reflect the real deployment (e.g. different cells/scenarios/ channel conditions) | Low, depends on training data | Low, depends on training data | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | High,  TE needs to cope with multiple decoders from multiple UE vendors | Relatively high,  TE needs to cope with multiple decoders from multiple NW vendors | Low,  TE needs to provide support for a limited number of test decoders that specified and captured in RAN4 spec. | Medium/Low,  TE needs to provide support for test decoders that specified(partially) and captured in RAN4 spec, training on TE may need. | | Specification Effort (e.g. test decoder) | Low,  Some conditions on the test decoder might be needed to ensure it can be implemented by TE.  Some assisted information on the test decoder might be needed to ensure the alignment between the test decoder and the encoder under test. | Low,  Some conditions on the test decoder might be needed to ensure it can be implemented by TE.  Some assisted information on the test decoder might be needed to ensure the alignment between the test decoder and the encoder under test. | High,  Consensus of a test model(s) in RAN4 is a challenging task. | Medium,  Less heavy workload than Option 3  Some assisted information on the test decoder might be needed to ensure the alignment between the test decoder and the encoder under test. | | Confidentiality/IP issues | Need to be considered | Need to be considered | Need to be considered | Need to be considered | | Applicability to different scenarios/conditions/ configurations | Depends on the granularity of test decoder and data set(s) | Depends on the granularity of test decoder and data set(s) | Depends on the granularity of test decoder and data set(s) | Depends on the granularity of test decoder and data set(s) | | Complexity of actual testing procedure for the ecosystem | High | High | Low | Low | | Friendly to STOA(state of the art) model test | Yes | Yes | No | [No] | | Whether model transfer/delivery is needed during the test procedure | Yes | Yes | No | No |   **Proposal 3: Dataset based on TR 38.901, e.g. UMa channel, UMi channel, CDL channel, “legacy approach”, should be considered in RAN4.**  **Proposal 4: Regarding the AI/ML capabilities, following aspects should be considered**  **- Definition of basic AI/ML capability and corresponding testing metrics**  **- Definition of different AI/ML capability levels and different testing metrics for different levels**  **- Dynamic AI/ML capabilities** |
| [**R4-2320186**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320186.zip) | Huawei,HiSilicon | ***Proposal 1:*** Table for description of 2-sided model testing options.   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Option 1 | Option 2 | Option 3 | Option 4 | | Clarification of options | | | | | | Source of the test decoder | 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) |  | | 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)~~ | * Transparent to RAN4 spec, no spec to ensure that it’s supportive | * Transparent to RAN4 spec, no spec to ensure that it’s supportive | Transparent to RAN4 spec, no spec to ensure that it’s supportive.  Note: Maybe supportive only if qualified training dataset is also specified | Transparent to RAN4 spec, no spec to ensure that it’s supportive.  Note: Maybe supportive only if qualified training dataset is also specified | | Test decoder verification procedure at TE and/or DUT | Need | Need | Need | Need | | Feasibility of test decoder verification procedure | No consensus | No consensus | No consensus | No consensus | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | It depends on training dataset | It depends on training dataset | It depends on training dataset | It depends on training dataset | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | It depends on the total number of the test model | It depends on the total number of the test model | It depends on the total number of the test model | It depends on the total number of the test model | | Specification Effort (e.g. test decoder) | * Procedure for verifying the decoder | * Procedure for verifying the decoder | * Align on assumptions for both for model structure and for model parameters * It also depends on whether model structure per use case or per configuration/scenario, and whether model parameters per configuration/scenario | Depend on which part of the test decoder is specified and which entity provides the unspecified part | | Confidentiality/IP issues | YES (if model exchange between TE vendor and UE vendor) | YES (if model exchange between TE vendor and NW vendor) | NO | NO | | ~~Applicability to different scenarios/conditions/ configurations~~ |  |  |  |  | | ~~Complexity of actual testing procedure for the ecosystem~~ |  |  |  |  |   ***Proposal 2: The interoperability is verified via core requirements and performance requirements.*** |
| [**R4-2320416**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320416.zip) | MediaTek inc. | In Table. 1, we provide our analysis/view on different options for the left parts.  Table 1. Summary of test decoder design options for 2-sided models   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Option 1 | Option 2 | Option 3 | Option 4 | | Clarification of options | | | | | | Source of the test decoder | 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) | Up to TE vendor. Alignment with DUT/infra vendors 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) | Type 1/2/3  Depending on DUT implementation | Type 1/2/3  Depending on cooperation between DUT and infra/TE vendors | Type 1/2/3  Depending on cooperation between DUT and TE vendors | Type 1/2/3  Depending on cooperation between DUT and TE vendors | | Test decoder verification procedure at TE and/or DUT | Not needed for DUT  Needed for TE to verify the compatibility | Not needed for DUT  Needed for TE to verify the compatibility | Not needed for DUT  Not needed for TE | Needed for DUT  Not needed for TE | | Feasibility of test decoder verification procedure | DUT side: NA  TE side: DUT needs to cooperate with TE | DUT side: infra vendor needs to cooperate with DUT  TE side: infra vendor needs to cooperate with TE | DUT side: NA  TE side: NA | DUT side: TE needs to cooperate with DUT  TE side: NA | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | The test decoder provided by DUT vendor may not reflect the actual decoder implemented by infra vendor | The test decoder provided by infra vender may reflect the performance in the field since infra vendors may use the same or similar decoder in the field as the test decoder | Depends on the test decoder decided during discussion | Depends on the test decoder decided during discussion | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | TE needs to support various test decoders from DUT vendor | TE needs to support various test decoders from infra vendor | Single implementation, TE develops the test decoder fully specified in RAN4 specification | Single implementation, TE develops the test decoder partially specified in RAN4 specification and also the unspecified part. | | Specification Effort (e.g. test decoder) | little effort  Need some limitation on test decoder to ensure it can be implemented on TE side | little effort  Need some limitation on test decoder to ensure it can be implemented on TE side | Consensus of a reference model in RAN4 may be a challenging task | Less heavy workload than fully specify test decoder in Option 3 | | Confidentiality/IP issues | Need to be considered. Model exposure is required from DUT to TE | Need to be considered. Model exposure is required from infra vender to TE | No issue | No issue | | Applicability to different scenarios/conditions/ configurations | Maybe | Maybe | Maybe | Maybe | | Complexity of actual testing procedure for the ecosystem | High  Need cooperation between DUT and TE | High  Need cooperation between infra vender and TE | Low  TE only needs to implement the test decoder | Low  TE only needs to implement the test decoder | |
| [**R4-2320555**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320555.zip) | ZTE Corporation | In this contribution, we have the following observations and proposals for the AI/ML :  **Observation 1: Level x is implementation-based AI/ML operation without any dedicated AI/ML-specific enhancement (e.g., LCM related signaling, RS) collaboration between network and UE and no collaboration for level x based on RAN1 outcomes.**  **Proposal 1: RAN4 shall not study the interoperability aspect for level x based on previous meetings in RAN1.**  **Observation 2: For the level y collaboration, it is clarified as the signaling-based collaboration without model transfer.**  **Proposal 2: RAN4 needs to consider the interoperability for collaboration level y based on more RAN2 progress since it is the signaling-based collaboration.**  **Observation 3: Two categories of models including Proprietary-format models and Open-format models were proposed by RAN1. For the Proprietary-format models, due to the lack of inter-operation and recognition between vendors, it is hard to standardize based on the unified specification identification.**  **Observation 4: Regarding to the Open-format models, the interoperability is feasible.**  **Proposal 3: RAN4 can focus on the Open-format models firstly, and discuss which core part and performance part requirements should be identified and how to define. On the other side, the test framework and procedure should also be discussed. At the meanwhile, RAN4 needs to wait for RAN1 progress on Open-format models.**  **Proposal 4: RAN4 shall study the basic structure of the open-format and consider how the common understanding defined between different vendors.**  **Proposal 5: From the perspective of test, both functionality test and performance test should be considered.**  **Observation 5: model inference is the core component of AI/ML. Two aspects should be considered to verify: 1) The outputs are the results from the AI/ML inference model rather than the traditional solution; 2) The accuracy of outputs meet the requirement.**  **Observation 6: the latency requirements and the KPI and the threshold for judging the model performance shall be considered in RAN4.**  **Observation 7: The test environment is quite different from the field, so the assumptions is a little limitations and the generalization will be degraded.**  **Observation 8: The options may not work well, and the requirements based on these assumptions may not have significance.**  **Observation 9: The TU is limited in RAN4 for AI/ML.**  **Proposal 6: RAN4 shall study one-sided model firstly and the discussion of two-sided model shall be deprioritized since the two-sided model is only for CSI case.**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **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 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) | Options:   1. RAN4 specifications FFS whether alignment with UE/gNB vendors is required, 2. Up to decoder implementer (TE vendor) FFS whether alignment with UE/gNB vendors is required 3. Combination of Option 1 and 2 | | 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)*.* | No need to consider training collaboration;  Type1/2/3 | No need to consider training collaboration;  Type1/2/3 | No need to consider training collaboration;  Type1/2/3 | No need to consider training collaboration;  Type1/2/3 | | Test decoder performance verification procedure at TE and/or DUT | Not needed (UE has full knowledge) | Needed (it needs to be proved that the decoder works such that test failure is because of DUT) | Not needed (fully specified in spec.) | Not needed (Test validation would follow the typical RAN5/testing procedures)  Needed | | Feasibility of test decoder verification procedure | NA | Feasible if network vendor provides a test encoder and passes the test  Other options? | NA | Feasible if TE vendors shares test decoder | | Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | [One] | [Option A: One  Option B: More than one  Option C: RAN4 doesn’t need to make decision] | [One] |  | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | Low   1. There could be large performance mismatch with field performance due to mismatch between test decoder and field decoder implemented by gNB 2. Depends on the training data set | High  Model which is similar with the test model could be used in the actual deployment by gNBs | Low  There could be large performance mismatch with field performance due to mismatch between test decoder and field decoder implemented by gNB | Medium/Low   1. The test decoder may have a large mismatch with the decoders deployed in the field, and UE may easily pass the test since UE could train the model based on the specified decoder 2. May partially reflect the performance in real deployment based on specified parts of test decoder 3. Depends on the data sets used for training | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | 1. TE will need to support a wide range of architectures/interfaces/algorithms (at least one per UE vendor). TE Computational resources requirements should be defined. No additional training required by TE vendor. 2. Medium. TE needs to cope with multiple decoders from multiple UE vendors | 1. TE will need to support a considerable range of architectures/interfaces/algorithms (at least one per infra vendor). TE Computational resources requirements should be defined. No additional training required by TE vendor. 2. Medium. TE need to implement multiple decoders from different BS vendors 3. Hgh | Low  Least complexity required on the TE side | Medium  If TE is responsible for the training of test decoder, the requirements for TE is high. If UE or gNB provide the test decoder, the requirements for TE is the same as option 1 and 2. | | Specification Effort (e.g. test decoder) | Low  Some conditions on the test decoder might be needed to ensure it can be implemented by TE | Low  Some conditions on the test decoder might be needed to ensure it can be implemented by TE | High  Long/complicated discussion expected in RAN4 to derive the test decoder | Medium to high  Less heavy workload than Option 3  Long/complicated discussion expected in RAN4 to agree on the test decoder to be specified in RAN4 | | Confidentiality/ IP issues | FFS  Need to be considered  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 | FFS  Need to be considered  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 | None  Decoder is fully captured in the specifications | 1. None 2. Depending on the source of data used for training the model to be specified, there might be confidentiality issues in this option. | | Applicability to different scenarios/conditions/ configurations | Maybe   1. pending on how to design the test to guarantee the generalization 2. The model should be applicable for the scenarios/ configurations tested for in RAN4 | Maybe   1. pending on how to design the test to guarantee the generalization 2. The model should be applicable for the scenarios/ configurations tested for in RAN4 | Maybe   1. pending on how to design the test to guarantee the generalization 2. The model should be applicable for the scenarios/ configurations tested for in RAN4 | Maybe   1. pending on how to design the test to guarantee the generalization 2. The model should be applicable for the scenarios/ configurations tested for in RAN4 | | Complexity of actual testing procedure for the ecosystem | Medium/High  Potentially, for each DUT, TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test. When executing test, DUT vendor will need to make a manufacturer declaration indicating the test decoder(s) they want to be tested against and for which scenarios (only one or more than one if the DUT is using different AI/ML models for different scenarios). | High  DUT will need to be tested against one or multiple test decoders provided by different NW vendors (manufacturer declaration?)   Potentially, for each NW test decoder (or even test decoder update?), TE vendor will need to integrate its test decoder (if not leveraging from a previous design) before enabling test. | Low  No additional TE integration required once initial implementation of the test system is completed. | Low/Medium  All DUT are supposed be tested against equivalent TE vendor implementation of the test decoder (only one).  No additional TE integration required once initial implementation of the test system is completed. | | Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes | Yes | No | No/Maybe | | Relationship with reference decoder/encoder for defining requirement | [Alt 1: same as reference decoder  May not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.  Alt 2: different from reference decoder  UE may not pass the tests due to different test decoders are used for defining requirements and tests.] | [Alt 1: same as reference decoder  May not be possible to define requirements as there could be larger performance gap among companies. The results may not be able to be calibrated.  Alt 2: different from reference decoder  UE may not pass the tests due to different test decoders are used for defining requirements and tests.] | [Alt 1: same as reference decoder  Possible to define requirements and be able to calibrate results from companies.  Alt 2: different from reference decoder  There is no reason to specify test decoder different from that is used for defining requirements. ] | [Alt 1: same as reference decoder  There is good chance that the results among companies can be calibrated as the performance of the model could largely be decided by the specified part.  Possible to define requirements  Alt 2: different from reference decoder  There is no reason to specify different test decoder than that is used for defining requirements.] | | Whether model transfer/delivery is needed during the test procedure | [Yes] | [Yes] | [No] | [No] |   **Observation 10: 3GPP channel models have stable performance and sufficient physical meanings. It is convenient to generate large number of samples using 3GPP channel models.**  **Proposal 7: In order to guarantee the stable performance and convenience, RAN4 shall study and use the dataset based on TR 38.901 firstly.** |
| [**R4-2320611**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_109/Docs/R4-2320611.zip) | Samsung | *Interoperability: Two-sided model framework*  **Proposal 1: For Option 4,**  **- It is expected/assumed that TE vendor will not share test decoder to other vendors (DUT vendors and/or infra vendors);**  **- Standardized dataset is required to guarantee other vendors to develop decoder for similar performance:**  **🡺 standardized dataset include: Target CSI and CSI feedback.**  **Proposal 2: The following clarification of options are provided for option 1-4 test decoder for 2-sided model.**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Option 1 | Option 2 | Option 3 | Option 4 | | Clarification of options | | | | | | Source of the test decoder | **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)** | Depends on TE vendors’ procedure for decoder development | | 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) | **Type 1**  (Joint training of encoder/decoder  at UE-sided) | **Not applicable** (if test decoder is not provided to UE vendors for encoder design)  **Or**  **Type 2 or Type 3 (NW first)**  (Type 2: Only if gradient results or test decoder can be provided to UE vendors for encoder design) | **Type 2 or Type 3 (NW first)** | **Maybe Type 3 (NW first)**  (if Type 3 collaboration procedure is specified and followed by TE/DUT vendors) | | Test decoder verification procedure at TE and/or DUT | **Maybe, but depends on verification procedure feasibility** | **Maybe, but depends on verification procedure feasibility** | **No**  (3GPP-specified test decoder leads to same TE implementation) | **No**  (3GPP-partially-specified test decoder and retuned by TE vendor) | | Feasibility of test decoder verification procedure | **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) | **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) | **Not applicable** | **Not applicable** | | Pros/Cons analysis | | | | | | Reflection on the real deployment (knowledge of model, training type, etc.) | **No**  (Can’t reflect real deployment since no evidence shown that BS vendors will adopt decoder provided by UE vendors) | **Yes or Maybe** (Depends on test decoder can be provided to UE for model design) | **Maybe** (Depends on whether specified test decoder can reflect decoder in the field) | **Maybe** (Depends on whether specified test decoder can reflect decoder in the field) | | TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | **Limited effort**  (Model complexity assumption should be aligned by TE and UE vendors) | **Limited effort**  (Model complexity assumption should be aligned by TE and BS vendors) | **No effort** | **No effort** | | Specification Effort (e.g. test decoder) | **No specification effort required for test decoder** | **No specification effort required for test decoder** | **RAN4 effort required on specifying test decoder**  (including model structure and dataset for training etc.) | **RAN4 effort required on specifying test decoder**  (including model structure and dataset for training etc.) | | Confidentiality/IP issues | **Yes** (Disclosure of UE vendor designed IP) | **Yes** (Disclosure of BS  vendor designed IP) | **No issues identified** | **Maybe**  (Disclosure of TE vendor designed IP, but depends on model retuning procedure) | | Applicability to different scenarios/conditions/ configurations | **Yes**, if UE vendors can provide different test decoders accordingly | **Yes**, if BS vendors can provide different test decoders accordingly | **Yes**, if 3GPP can specify different test decoders accordingly | **Yes**, if 3GPP can specify different test decoders and/or TE vendor retune the model accordingly | | Complexity of actual testing procedure for the ecosystem | **Low** (DUT is only required to be tested against the specified test decoder) | **High** (Conformance tests could be not available since different test decoders from different BS vendors) | **Low** (DUT is only required to be tested against the specified test decoder) | **Low/Medium** (DUT is only required to be tested against the partially specified test decoder, but refinement procedure maybe required for TE) |   **Observation 1: For the reference decoder for test implementation for two-sided models in the UE performance tests, the feasibility of the offline training to obtain UE encoder can be confirmed at least for Option 1 and 3.**  **Observation 2: Only Type-1 and Type-3 training collaboration with the offline training manner needs to be considered in Rel-18 RAN4 study on the methodology to obtain the test model for two-sided model test implementation.**  **Observation 3: For the reference decoder to be used in the test implementation for two-sided models for the UE performance tests:**  **- Option 1 can be regarded to match with Type-1 training collaboration, i.e., decoder developed by UE vendors shall be provided to and used by BS vendors directly;**  **- Option 2 (if test decoder can be provided to UE for model design) can be regarded to match with partially Type-3 training collaboration, i.e., decoder is provided by gNB vendors for UE-side training, but without further gNB-side training based on labeled data.**  **- Option 3 can be regarded to match with partially Type-3 training collaboration, i.e., decoder is provided UE-side training, but without further gNB-side training based on labeled data.**  **- Option 4 can be regarded to match with Type-3 training collaboration, if the procedure of Type 3 collaboration can be followed by TE and UE vendors.**  **Proposal 3: In Rel-18 study item, RAN4 shall not narrow down any of four options, while only study and capture pros/cons in TR.**  *Interoperability aspects except 2-sided model issues*  **Proposal 4: RAN4 capture the below table in TR, which is the interoperability aspects for different level of NW-UE collaboration.**  **Table: Interoperability aspects for different level of NW-UE collaboration**   |  |  |  |  | | --- | --- | --- | --- | |  | Model Training | Model monitoring and Model selection/(de)activation/ switching/fallback | Model Inference | | N/W-UE Collaboration  Level-x | N/A (training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | N/A | Interoperability guaranteed by  - Use case KPI | | N/W-UE Collaboration  Level-y | N/A (training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.) | Interoperability guaranteed by  - Model monitoring perf.  - Model selection/(de)activation/ switching/fallback perf. | Interoperability guaranteed by  - Use case KPI | | N/W-UE Collaboration  Level-z | N/A for one-sided model training (training in non-3GPP entities or offline training as baseline, model training perf. guaranteed by model inference perf.)  N/A for two-sided model online training and FFS offline training. | Interoperability guaranteed by  - Model monitoring perf.  - Model selection/(de)activation/ switching/fallback perf.  No interoperability aspects for   - model deployment /update/transfer/delivery from/to model storage | Interoperability guaranteed by  - Use case KPI |   *Testability: Reference block diagrams for testing*  **Proposal 5: RAN4 shall firstly discuss and agree on the following principles to draft the reference block diagram for 1-sided model and 2-sided model:**  **- shall NOT contain the block for training;**  **- shall contain the blocks for model/functionality monitoring and selection/switching/ (de)activation/ fallback in DUT;**  **- shall contain the AI/ML LCM procedure verification and model control in TE;**  **- shall contain the test scenario generator to enable testing in different scenarios, used for generalization verification aspects.**  **Proposal 6: The purpose of introducing the diagram (to be captured in TR38.843) is to derive the potential testing procedure and used as the basis to judge whether certain performance metric is testable, for each use case for normative work.**  **Proposal 7: RAN4 shall include the following reference block diagram in TR for testing 1-sided model (UE as DUT).**    **Fig 2: Reference block diagram for testing 1-sided model (UE as DUT)**  **Proposal 8: RAN4 shall include the following reference block diagram in TR for testing the UE-side model of the 2-sided model (based on the example use case of CSI compression).**    **Fig. 3: Reference block diagram for testing UE-side model of the 2-sided model  (based on the example use case of CSI compression)**  **Proposal 9: Before defining reference block diagram for testing gNB-side model of the 2-sided model, the test metric and procedure shall be clarified for feasibility.**  **Proposal 10: FFS the feasibility of using NR air interface to test either 1-sided model implemented in gNB side or gNB-side model of 2-sided model. If not confirmed, gNB-side model shall be precluded for testing in RAN4.**  *Testability: Test data generation*  **Observation 4: Pros and cons for Option-a (dataset provided by 3GPP) and Option-c (methodology provided by 3GPP) are observed:**  **- Option-a: Pro: the dataset is provided so the training/test can be conducted accordingly. Cons: 3GPP have not yet provided a dataset for testing before.**  **- Option-c: Pros: no need to specify dataset; Cons: Even certain rules/function are specified, the dataset generated/used by TE vendors may be not large enough to cover all randomness due to test limitation (e.g., limited test duration), which can be a problem for repeatability of conformance testing.**  **Proposal 11: RAN4 adopt Option-c (TE generates data for test based on assumptions/parameters defined by RAN4 (e.g. by defining some rules/function to generate data)) as the baseline in normative phase:**  **- The testability (especially the repeatability of conformance testing) shall be further checked based on particular use case.** |

## Open issues summary

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

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 priority:

1. Reference block diagram for 1 sided model
2. Reference block diagram for 2-sided model
3. Option 4 clarifications
4. Test encoder/decoder options comparison discussion
5. Feasibility of different options

2nd priority:

1. Interoperability aspects

### Sub-topic 3-1

*Reference block diagram for 1 sided model*

There are 5 proposals for the block diagrams, it is proposed to start the discussion based on the simplest model provided (R4-2318935) and make further changes to converge on a diagram to be included in the TR.

Data collection

M/F management

Verification

AI/ML modules

AI/ML modules

Test configuration/controller

Test data generator

Model inference

M/F management

M/F monitoring

DUT

TE

**Issue 3-1: Block diagram for 1-sided model**

* Proposals
  + Option 1: Add description of each box (some proposals contained descriptions inside the boxes)
  + **Model inference (R4-2306299)**

***…. AI/ML based performance enhancements mainly focus on how to define requirements and tests for inference***

* + **Functionality/Model(M/F) monitoring procedure (R4-2306299)**

**(Study) *Performance (Model/Functionality) monitoring procedure, including performance evaluation and decision-making procedure …***

* + **Functionality/Model(M/F) management procedure (R4-2306299)**

**(Study) *Functionality/Model management procedure, including functionality/model selection/activation/deactivation, and functionality/model switching/fallback/transfer/delivery/update***

* + **Test data generator (R4-2306299)**

**(Study) *Different generating methods of test dataset can be used for different tests. The following candidate methods are to be considered …***

* + Option 2: Add other boxes in the TE or DUT
  + Option 3: Other additions/changes
* Recommended WF

Discuss and agree changes

### Sub-topic 3-2

*Reference block diagram for 2-sided model*

There are 5 proposals for the block diagrams, it is proposed to start the discussion based on the simplest model provided (R4-2318935) and make further changes to converge on a diagram to be included in the TR.

Data collection

M/F management

Verification

AI/ML modules

AI/ML modules

Test configuration/controller

Test data generator

Model inference

M/F management

Model inference

M/F monitoring

DUT

TE

**Issue 3-2: Block diagram for 2-sided model**

* Proposals
  + Option 1: Add description of each box (some proposals contained descriptions inside the boxes)
  + **Model inference (R4-2306299)**

***…. AI/ML based performance enhancements mainly focus on how to define requirements and tests for inference***

* + **Functionality/Model(M/F) monitoring procedure (R4-2306299)**

**(Study) *Performance (Model/Functionality) monitoring procedure, including performance evaluation and decision-making procedure …***

* + **Functionality/Model(M/F) management procedure (R4-2306299)**

**(Study) *Functionality/Model management procedure, including functionality/model selection/activation/deactivation, and functionality/model switching/fallback/transfer/delivery/update***

* + **Test data generator (R4-2306299)**

**(Study) *Different generating methods of test dataset can be used for different tests. The following candidate methods are to be considered …***

* + Option 2: Add other boxes in the TE or DUT
  + Option 3: Other additions/changes
* Recommended WF
  + To be discussed

### Sub-topic 3-3

*Option 4 clarifications*

Several companies discussed Option 4 and are seeking further clarification

**Issue 3-3: Option 4**

* Proposals
  + Option 1: Interoperability should be ensured based on the specified parameters
  + Option 2: Partially standardized dataset for training of the test cased should be considered, further discuss in WI
    - Option 2a: ): Capture dataset with (nominal encoder input, latent message as encoder output) in the specification for TE vendors to train the decoder, to ensure test repeatability and ability for other vendors to train the decoder with similar performance
  + Option 3: Parameters to be specified
    - Model structure
    - Activation function
    - Maximum FLOPs allowed for the test decoder
    - Maximum number/size of parameters
    - Others
  + Option 4: Performance parameters
    - Cosine similarity threshold
    - γthreshold value )see R4-2318764
* Recommended WF
  + To be discussed, several options can be agreed

### Sub-topic 3-4

*Test encoder/decoder options comparison*

A table summarizing the comparison of the 4 options under study for the testing of the two sided model has been discussed in the past 2 meetings. In the previous meeting some agreements were reached and the agreed part is listed below for convenience (R4-2317631)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 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) |  |
| 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 |

Several companies submitted analysis with the rest of the table contents, these are discussed below. The table is the moderator’s proposal based on the companies’ inputs.

**Issue 3-4: 2-sided testing options comparison table**

* Proposals
  + Option 1:further discuss the table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 |  |  |  |  |
| Source of decoder training data |  |  |  | FFS  Could be specified depending on how Option 4 will be defined |
| DUT vendor knowledge of the test decoder |  |  |  |  |
| Supported training collaboration type between DUT and decoder provider (source of training data should be consistent with the collaboration type) | N/A  Training entirely up to DUT vendor | No need to discuss  Type 1/3 with NW first training  Type 2 offline | N/A | No need to discuss  Type 3 – depending on how decoder is specified |
| Test decoder performance verification procedure at TE and/or DUT | Needed | Needed | Not needed | Not needed  Needed |
| Feasibility of test decoder verification procedure | FFS | FFS | N/A | N/A |
| Number of test per test configuration/setup (propagation condition, CSI configuration etc excluding decoder/network side model configuration) | No need to discuss/Depends on the test | | | |
| **Pros/Cons analysis** | | | | |
| Reflection on the real deployment (knowledge of model, training type, etc.) | Low  Impossible to know whether it reflects any gNB implemented decoder | High | Low/Medium  depends on training data used when decoder is defined | Low/Medium  depends on training data used when decoder is defined |
| TE requirements to deploy the decoder (e.g. training, complexity, interoperability) | Medium  High – TE needs to support multiple test decoders | Medium  High – TE needs to support multiple test decoders | Low | High  Lowest – TE needs to support single test decoder  Medium – TE might have to support retraining/finetuning when models are updated |
| Specification Effort (e.g. test decoder) | Low | Low | High | Medium |
| Confidentiality/ IP issues | Potentially 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 | Potentially 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 | It depends on generalization test procedure  Yes | It depends on generalization test procedure | It depends on generalization test procedure  Conditional yes | It depends on generalization test procedure  Yes |
| Complexity of actual testing procedure for the ecosystem | Medium  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 | Low |
| Friendly to STOA(state of the art) model test / Forward compatibility when new AI models are invented | Yes | Yes | No | No/Maybe |
| Relationship with reference decoder/encoder for defining requirement | A different reference decoder (e.g., based on option 3 or option 4) for defining requirements.  Relationship between decoder/encoder used to define requirements and one used for testing is unknown | A different reference decoder (e.g., based on option 3 or option 4) for defining requirements.  Relationship between decoder/encoder used to define requirements and one used for testing is unknown | Same reference decoder as test decoder for defining requirements. | Same reference decoder as test decoder, or different reference decoder based on option 3 for defining requirements. |
| Whether model transfer/delivery is needed during the test procedure | Maybe – FFS how model is provided | No  Model cannot be provided during the test | No | No |

* Recommended WF
  + To be discussed

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

### Sub-topic 3-5

*Feasibility of different testing options*

At least a preliminary conclusion on feasibility of the testing options would be useful as a conclusion of the SI

**Issue 3-5: Feasibility of different testing options for 2-sided models**

* Proposals
  + Option 1: At least Option 3 is feasible
  + Option 2: At least Option 3 and 4 are feasible
  + Option 3: All options are feasible
  + Option 4: All options require more study
  + Option 5: Other
* Recommended WF
  + To be discussed

### Sub-topic 3-6

*Interoperability aspects*

**Issue 3-6: Interoperability aspects**

* Proposals
  + Option 1: Capture the following table in the TR:

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

* Recommended WF
  + Discuss whether this should be captured in the TR and what changes are needed, if any