**3GPP TSG-RAN WG1 Meeting #115 R4-2412831**

**Maastricht, Netherlands, 19th – 23rd August, 2024**

**Agenda item:** 8.17.5

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [112][129] 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 R4-2410570. This summary is organized in 4 high level topics and contains several sub-topics for discussion.

# Topic #1: General aspects

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2411258**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411258.zip) | CAICT | **Proposal 1: Consider effective performance monitoring for post-deployment handling of minor model update or fine-tuning.**  **Proposal 2: Whether and how to define test on model transfer/update shall wait for more explicit progress from other WGs.**  **Proposal 3: Suggest to discuss potential solutions for offline performance tests conducted inside vendor.** |
| [**R4-2411340**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411340.zip) | CATT | This paper discussed the general issues related to AI/ML for NR air interface, and following observations and proposals are provided:  **Observation 1: For a given dataset, at least hundreds of MBs is required for storage.**  **Observation 2: Dozens to hundreds of GBs are required for dataset storage every time when all datasets are updated.**  **Observation 3: Less storage size is required if platform files are shared, e.g., Pytorch/tensorflow files that include model structure and parameters.**  **Proposal 1: Only the datasets that are agreed to be put into the specifications are stored for a long term. Those for calibrations and discussions are stored temporarily.**  **Proposal 2: Large bandwidth may be needed for efficient download/upload, and concurrent download/upload may appear, which requires more study from MCC.**  **Proposal 3: Information in a readme file about the stored datasets at least includes:**   * **Scenarios** * **Configurations** * **Dataset format** * **Data type.**   **Proposal 4: A brief description of the AI/ML model functions should be stored together with the AI/ML models.**  **Proposal 5: The tools for storage should be capable of checking the validation of datasets and models, e.g., the following issues:**   * **Dimensions mismatch.** * **Invalid values, e.g., NAN, Inf.** * **Performance of AI/ML models, e.g., SGCS, is lower than a given threshold.** * **Others.**   **Proposal 6: RAN4 further check the demands on versioning of datasets/models and not to preclude this function at current stage.**  **Proposal 7: More efforts are required to study the schemes of AI/ML model sharing among different platform. And the impacts of the following issues may need study:**   * **Resolution loss.** * **Some operations are not supported by intermediate tools, e.g., ONNX.** * **Compatibility/version mismatch.** |
| [**R4-2411408**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411408.zip) | Apple | **Observation 1: "Model as a baseline" can either be explicitly captured in RAN4 specifications or agreed upon for aligning performance results.**  **Observation 2: For Option 1 and for verifying DUT’s AI/ML capability to load and execute the models, trained AI/ML models can be defined in RAN4 spec for different use case tests**  **Observation 3: For Option 1 and for ensuring the model is properly conducted, performance requirements can be established for RAN4-defined AI/ML models across various use cases. The DUT is considered to have successfully passed the tests if it meets the specified performance requirements.**  **Observation 4:** **The identified scenarios and configurations can be initially understood as those reported by UE through capability signaling** **as part of functionality identification.**  **Observation 5: Different scenarios that will be part of generalization test could act as the additional conditions for the AI/ML model training but do not constitute a part of UE capability for the AI/ML-enabled feature/FG**  **Observation 6: Configurations utilized for generalization test should be associated with UE capability of an AI/ML-enabled Feature/FG (set the configuration and vary the conditions under the configuration). Different conditions should be applied for each configuration.**  **Observation 7: The existence of a wide range of diverse UE capabilities poses a challenge for RAN4 in identifying a typical configuration or scenario for specifying test cases.**  **Observation 8: On device fine-tuning based on retraining could be beneficial to model delivery/transfer to reduce overhead/latency**  **Observation 9:** **Having a separate AI/ML model for each unique scenario, configuration, and additional condition could significantly increase the complexity and storage requirements of the User Equipment (UE). Additionally, it may introduce overhead in terms of model delivery, transfers, and associated latency**  **Observation 10: The post-deployment phase can be considered within the broader context of the generalization framework**  **Observation 11: Options 1-a and 1-b for post deployment testing would impose a large burden for testing, signal overhead and complexity.**  ***RAN4 performance testing goals***  **Proposal 1: Deprioritize Option 1 and focus on Option 2 for performance requirements as a testing goal**  **Proposal 2: For verifying performance gain of AI/ML models/functionalities, RAN4 can define multiple independent** **test cases with different scenarios and configurations/conditions as reported through UE capability signaling, which could include:**  **Propagation conditions/scenarios, e.g., channel modes defined for different scenarios (CDL, AWGN, etc.) in TR38.901, Doppler conditions, SNR levels, various deployment scenarios (Uma, Umi, ISD), different cell with different gNB height, various indoor/outdoor UE distributions, various UE mobility etc.**  **- Configurations:**   * **Various UE parameters: number of UE beams, antenna panels, antenna array dimensions, different AI/ML configurations: different set B of beams, T1 for measurement/T2 for prediction for BM-case2** * **Various gNB settings: Tx codebooks (set A/B beams), beam widths, antenna spacings, Tx antenna dimension, antenna port layouts, TXRU antenna virtualization (e.g. (8,8,2,1,1,2,8), vs (8,4,2,1,1,2,4))**   **RAN4 should conduct a down-selection process to narrow down the scenarios and configurations (parameters) for the conformance test. This test will serve as a benchmark for evaluating the model's performance across a range of unseen scenarios and configurations.**  ***Static/non-static scenarios/configurations***  **Proposal 3:** **If non-static scenarios/configurations are supported for certain use cases, they can be included as part of generalization tests.**  **Proposal 4:** **Non-static scenarios/configurations should be considered for test cases only if static scenario/configuration testing fails to fulfill the testing objectives. CSI and beam management temporal prediction use cases are particularly suitable for introducing non-static environments during testing**  **2.3 Generalization/scalability aspects**  **Proposal 5: To determine the granularity of additional scenarios/conditions for defining generalization tests for each use case, it's essential to study how the AI model's behavior changes with different scenarios and conditions.**  **Proposal 6: To improve the generalization behavior of the model, training with a diverse dataset should be investigated.**  **Proposal 7: Investigate the reduction of generalization tests by training with a mixed dataset containing samples from different configurations. Investigate the definition of a single generalization test where the testing data consist of a random mixture of these configurations. If the performance degradation between the model trained on the mixed dataset and tested with random mixture configurations and the model trained and tested specifically for each configuration meets a predefined performance margin criterion, the test could pass.**  Proposal 8: Study post-deployment procedures to augment conformance testing for effectively managing performance across all possible deployment conditions/scenarios (which are not tested)  **Proposal 9: For UE-side models and/or UE-part of two-sided models** **it is suggested that the scenarios and configurations utilized for generalization tests can be determined based on the supported configuration reported by the UE as part of capability signaling.**  **Proposal 10: In the RAN4 core requirement, it is mandated that the consistency or association between of additional conditions during both training and inference is guaranteed. This serves as additional assistance information from the network side for testing UE-sided models during inference and monitoring**  **Proposal 11: RAN4 will explore methods to specify the identified scenarios and/or configurations per use case in future release, contingent upon other WGs can specify the granularity and the capability signaling.**  **Proposal 12: For defining generalization tests, RAN4 should define identified scenarios associated with the UE capability report of an AI/ML-enabled Feature FG, and other scenarios (additional conditions). RAN4 should also define minimum level of performance for the identified scenarios and/or conditions.**  **Proposal 13: Other scenarios and/or configurations can be interpreted as the scenarios and/or configurations that are not reported by UE capability signaling for an AI/ML-specific functionality or model ID.**  ***Scenario/configuration specific Models (Fine-tuning)***  **Proposal 14: RAN4 to discuss the practicality of formulating a framework that facilitates on-device fine-tuning. The focus will be on exploring the feasibility of creating a dynamic and site-specific approach to online training and fine-tuning (e.g reinforcement learning)**  **Proposal 15: UE to update its stored AI/ML models with the new model ID after fine-tuning, where this ID can be associated with the training data (which implicitly have the additional conditions) used to fine tune the model.**  Proposal 16: RAN4 should investigate the options for enhancing the generalizability of AI/ML models by providing the appropriate assistance/side information as input signal to the inference engine of the AI/ML model and discuss the feasibility of training with diverse datasets across different additional conditions  ***Principles on the definition of requirements***  Proposal 17: RAN4 should clarify/agree that the side conditions of the testing procedures should remain the same for legacy and AI/ML methods.  ***Requirements for LCM (Performance Monitoring)***  **Proposal 18: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case**  **Proposal 19: RAN4 shall consider the latency requirements for model monitoring input data as well as the establishment of tolerance margin requirements for the specified KPIs for model monitoring per use case**  **Proposal 20: FFS on how to perform cell level BM performance monitoring when the AI/ML model resides at NW**  ***Post Deployment validation and Fine tuning***  **Proposal 21: If the AI/ML model drifts due to misalignment of network-side additional conditions, the alignment could be achieved through an ID assigned by the network during training data collection. This ID indicates the association of the training data with the additional conditions implied to generate those data**  **Proposal 22: If the AI/ML model drifts due to misalignment of network-side additional conditions, if the UE lacks the ID but the network possesses the model ID, then the network can transfer the model to the UE, and the UE can update its list of models**  **Proposal 23: If the AI/ML model drifts due to misalignment of network-side additional conditions, if the UE has a model that partially supports the network-sided additional conditions, then the network can trigger data collection for fine-tuning the UE-sided model. Subsequently UE updates its list of model IDs.**  **Proposal 24: The UE updates its model list with a new model derived from either fine-tuning, model transfer from the NW, or monitoring to ensure the consistency of additional conditions. Then, the UE assigns an ID to the new model that supports the NW's additional conditions and shares this information with the NW. If some of the new conditions/configurations are standardized, the UE updates its capability report accordingly.**  **Proposal 25: If there is no ID available to associate training data with additional conditions, and monitoring procedure fails to guarantee the consistency of the model with the additional conditions then UE should fallback to legacy mode.**  ***Post-Deployment***  **Proposal 26: Study an on-device model monitoring/performance assessment to monitor active and inactive AI/ML models with the actual UE hardware and field data to dynamically manage the database of models and enable proactive monitoring for computing/updating KPIs and facilitates seamless transitions to newly updated or introduced models when needed, ensuring efficient model management.**  **Proposal 27: RAN4 to employ the post-deployment procedures described through the dynamic management of an AI/ML model database and RRC signaling for enabling monitoring procedures**  A diagram of a model  Description automatically generated  **Fig. 6:** **Post-deployment procedure through the dynamic management of an AI/ML model database and RRC signaling for enabling monitoring**  A diagram of a company  Description automatically generated  **Fig. 7:** **Post-deployment procedure**  ***Data collection for testing***  **Proposal 28:** **RAN4 must conduct an analysis for each use case to determine the reliability of using synthetic channels for test data in evaluating models trained on real data.** |
| [**R4-2411625**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411625.zip) | Xiaomi | **Observation 1:** **The multitude of options for post-deployment arises from the variety of cases and scenarios associated with model updates.**  **Proposal 1:** **RAN4 may categorize model updates based on different cases and evaluate whether a unified approach or multiple solutions are required for post-deployment.**  **Proposal 2: RAN4 to identify difference between conformance test and performance monitoring. It is advisable for RAN4 to consider performance monitoring under conditions encountered after post-deployment when defining conformance test requirements.** |
| [**R4-2411980**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411980.zip) | CMCC | ***Proposal 1: for post deployment handling, the option that RAN4 should discuss a framework to ensure that at least a validated model exists at the UE and/or the option that RAN4 should discuss a framework to ensure that non-validated/non-tested models are not used by the UE can be considered as high level principle, and further details can be FFS.***  ***Proposal 2: for UE-side models and/or UE-part of two-sided models, it is proposed that the scenarios and/or configurations used for generalization can be decided based on the supported configuration reported by UE.***  ***Proposal 3: for generalization, it is proposed to take the requirements for inference as the minimum level performance for generlazation.*** |
| [**R4-2412021**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412021.zip) | Nokia | **On Testing goals:**   1. For both Options of the testing goals, the criterion for passing the test is achieving of minimal performance/core requirements defined for the functionality, i.e., it is not feasible to define requirements per each identifiable model. 2. RAN4 needs to clarify whether the requirement/test defined for a specific model (Option 1 in testing goals) is applicable for the functionality. 3. RAN4 to agree that the requirements shall be defined on functionality level (possibly, different requirements in different configurations/parameters) even when a specific/reference model is used to derive the requirements. 4. Regarding FFS on the specific model: companies can select any specific models to pass the test (it can be reference model or not) but this model should be available at the device after deployment (e.g., for the fallback). 5. Testing goals and derivation of RAN4 requirements are two different aspects. Formulation of the requirements can be based on a certain model that is used to derive minimal level of performance for the whole functionality. Whereas, testing of each individual model will be much more challenging or nearly impossible for Release 19 due to lack of support from RAN1 and RAN2 mechanisms. 6. RAN4 to consider the following unified definition for the testing goal: The testing goal is to verify whether the minimum performance of AI/ML functionality/feature can be achieved in different test configurations/parameters.   **On Post-deployment handling:**  Current Option 2 based on monitoring does not completely address the issue of the use of the models/functionalities in the live network, that have not passed conformance testing but can demonstrate the performance below minimal requirements.   1. RAN4 to ensure (and to clarify in Option2: performance monitoring-based post-deployment handling) that at least one functionality (for an AI/ML-enabled Feature) that has passed conformance tests shall be always present in the device for the fall-back. 2. RAN4 to consider handling post-deployment verification of updated/new functionality in inactive/standby state/mode before it is allowed to be used along with the inference of the currently used functionality. 3. RAN4 to send an LS to RAN2 and RAN1 WGs on a need to consider conformance testability of AI/ML enabled functionalities when designing such LCM operations as monitoring/activation/update/fallback for the devices already deployed in the field.   **On AI/ML logistics:**  The means/tools for reference AI/ML models and/or datasets logistics could serve either ongoing RAN4 study where companies need to exchange own data or, in a longer term, to store standardized reference models/datasets, if those are needed and agreed for a specific use case.   1. RAN4 to focus only on the discussion of AI/ML logistics for the ongoing RAN4 study/experiment, i.e., do not consider any reference models/datasets that could be a part of the standard. 2. Reference AI/ML model sharing could be arranged even using exiting FTP due to their reasonable size, e.g., using Other TDoc type. 3. RAN4 to limit the possible formats of the shared AI/ML models to models developed in popular frameworks (e.g., TensorFlow and PyTorch) with support for intermediary sharing platforms (e.g., ONNX). 4. For dataset sharing, RAN4 needs to discuss further a new dedicated solution together with MCC considering the following aspects:    1. The size of the dataset on the level of 10GB per company per meeting,    2. Access to the datasets, e.g., only based on ETSI account,    3. Retention/history/optimization of the dataset storage, e.g., it can be limited to certain time horizon (1 year) or only to certain limited versions of the datasets,    4. Format of the dataset, e.g., limited to some of the popular open formats, such as mat or parquet. |
| [**R4-2412023**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412023.zip) | Nokia | CR for 38.843 with the following cover sheet:   |  |  | | --- | --- | | ***Reason for change:*** | Introduced definitions are not applicable to conformace testing. | |  |  | | ***Summary of change:*** | Added a note to the definitions of the following terms:   * AI/ML model testing * AI/ML model validation | |  |  | | ***Consequences if not approved:*** | The definitions of the terms can be considred generic and one can confuse those with conformace testing practices. | |
| [**R4-2412249**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412249.zip) | vivo | **Proposal 1: RAN4 to revisit the RRM requirements scope for LCM procedures based on the following agreement when RAN1/RAN2 have enough inputs for LCM procedures**   * ***RAN4 considers defining RRM requirements for LCM procedures based on the specific use cases (beam management and positioning), and candidates are:*** * ***identification*** * ***selection*** * ***activation*** * ***deactivation*** * ***switching*** * ***fallback to non-AI operation*** * ***performance monitoring*** * ***Others are not precluded*** |
| [**R4-2412331**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412331.zip) | Intel Corporation | **Proposal #1: Consider the following framework for post-deployment model (feature/functionality) verification:**   * + **UE shall support at least some default/baseline AI/ML models/features that passed conformance requirements. The respective models can be used as a fallback mechanism.**   + **RAN4 to support both approaches including conformance testing and performance monitoring. Further discuss conditions/scenario when each approach is applied.**   + **Changes or updates to the AI/ML models/features are tested by the device vendors against RAN4 requirements before the deployment to the UE is performed. The information on whether AI/ML model/feature update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**   + **Further discuss a potential specific performance validation procedure during AI/ML model transfer/update as a part of performance monitoring framework (involving RAN1).**   **Proposal #2: Define reference AI/ML models for one-sided and two-sided models (both encoder/decoder) for performance requirements definition taking into account performance/complexity tradeoffs.**  **Proposal #3: Postpone discussion on LCM requirements to Q4’24 and handle it in the RRM session.**  **Proposal #4: Adopt the following channel modelling framework for AI-ML requirements**   * + **Use synthetic channels as the default assumption for both training and inference phases. Assess their suitability for each specific use case.**   + **Field datasets should only be used if significant issues are identified with synthetic models.**   + **For training AI-ML models use generic synthetic channel model datasets based on TR 38.901, e.g. UMa channel, UMi channel, CDL channel, “legacy approach”, etc.**   + **For the inference (i.e., testing phase) use a simplified set of synthetic channel models, such as TDL (Tapped Delay Line) models or other channels used in legacy RAN4 tests.** |
| [**R4-2412765**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412765.zip) | Huawei,HiSilicon | ***Proposal 1:*** RAN4 test principle of AI is that DUT should not only meet but also exceed legacy non-AI requirements in AI test.  ***Observation 1:*** For some cases, to verify AI performance, new test set up is under discussion.  ***Proposal 2:*** If new test set up is introduced for AI test, two options are identified for defining AI performance requirement.   * Option 1: Define non-AI requirements under new test set up subject to AI. Use the non-AI requirement to define AI requirement, where the AI requirement has to be higher than that of non-AI.   + FFS: Feasibility of aligning non-AI performance baseline under new test set up * Option 2: Define absolute requirement of AI without considering the comparison with non-AI.   + FFS: how to reflect the AI test principle that DUT should not only meet but also exceed legacy non-AI requirements in AI test   ***Proposal 3:*** For generalization test, other scenarios and/or configurations are the scenarios and/or configurations that are not reported as applicable by UE according to RAN2 latest progress.  ***Proposal 4:*** For generalization test, performance for other scenarios and/or configurations can be ensured by RAN4 legacy non-AI test. |
| [**R4-2413040**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413040.zip) | ZTECorporation,Sanechips | **Proposal 1: 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 2: The minimum level of performance shall be studied per use case.**  **Proposal 3: 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 4: 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**.  **Observation 1: 1. The different use case has the different synthetic channels. 2. One synthetic channels shall be used for the individual use case**  **Proposal 5: RAN4 shall confirm that the different use case has the different synthetic channels.**  **Proposal 6: RAN shall consider how to design the synthetic testing data.** |
| [**R4-2413391**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413391.zip) | Ericsson | * ***Proposal 1:*** *Further clarify the options in the earlier RAN4 agreement as follows (the updates are in red):* * *To ensure the AI performance after device deployment, discuss the following options further*   + *Option 1: Conduct the ~~conformance~~ testing for new/updated AI model/functionality before its deployment in already deployed UEs (considering the UE hardware that it will be used in)*     - *FFS on the feasibility*   + *Option 2: Design the test to verify the performance monitoring and proactive recovery from potential performance degradation*     - *Depend on the other WG progress*     - *Monitoring can be used for managing fallback, model update/model switching/model transfer, if applicable*   + *Other options are not precluded* * *Proposal 2: As a further option relating to post deployment testing, consider the possibility of capturing model input during testing for later testing of new models. Add Option 3 to the earlier agreed Option 1 and Option 2:*   + *Option 3: Capture model input during conformance testing for later testing of new models.* * *Proposal 3: When discussing the feasibility of Option 1, RAN4 to consider its scalability limitation and potential mismatch between testing and actual environment.* * *Proposal 4: RAN4 to study and compare the feasibility of Options 1, 2, and 3. RAN4 can inform other groups about the conclusions of this study.* * *Proposal 5:* *RAN4 will study a maximum delay requirement for switching between AI/ML and non-AI/ML operation modes, at least in the fallback direction (AI/ML →non-AI/ML).* * *Proposal 6: The maximum fallback delay may include one or more components, e.g.,*   + *obtaining and starting to use a fallback tx/rx configuration,*   + *performing the necessary transmission and/or detection, identification, or measurement procedures,*   + *informing the network about the fallback decision (e.g., if the UE decides this autonomously) or about completion of the fallback, etc.* |

## Open issues summary

The open issues were grouped in the following sub-topics for further discussion based on the input contributions.

1. Post deployment testing
2. Synthetic data vs. field data
3. Testing goals
4. AI/ML performance
5. Generalization
6. CR from Nokia
7. LCM handling

### Sub-topic 1-1

*Post Deployment testing*

The need for post deployment testing (e.g. after a model is updated) was discussed in previous meetings. A high level agreement was reached in RAN4#110 and is reproduced below.

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

**Agreement:**

* To ensure the AI performance after device deployment, discuss the following options further
  + Option 1: Conduct the conformance testing for AI model/functionality before deployment
    - FFS on the feasibility
  + Option 2: Design the test to verify the performance monitoring
    - Depend on the other WG progress
    - Monitoring can be used for managing fallback, model update/model switching/model transfer, if applicable
  + Other options are not precluded

There was further discussion in RAN4#110Bis and RAN4#111 but there were no further concrete agreements. Conclusion from the discussion in RAN4#111 was that interested companies should bring more concrete analysis on the feasibility of the options discussed so far. The contributions in this meeting do not present much analysis on the feasibility of the options.

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

* Proposals
  + Option 1: Amend the agreements above as below (R4-2413391):
  + *Option 1: Conduct the ~~conformance~~ testing for new/updated AI model/functionality before its deployment in already deployed UEs (considering the UE hardware that it will be used in)*
    - *FFS on the feasibility*
  + *Option 2: Design the test to verify the performance monitoring and proactive recovery from potential performance degradation*
    - *Depend on the other WG progress*
    - *Monitoring can be used for managing fallback, model update/model switching/model transfer, if applicable*
  + *Other options are not precluded*
  + Option 2: Add Option 3 as below (R4-2413391):
  + *Option 3: Capture model input during conformance testing for later testing of new models.*
  + Option 3: UE shall support at least some default/baseline AI/ML models/features that passed conformance requirements. The respective models can be used as a fallback mechanism. (R4-2411980, R4-2412021 & R4-2412331)
  + Option 4: Drop Option 1 as framework is not clear and burden on UE vendors is very high
  + Option 5: RAN4 to consider handling post-deployment verification of updated/new functionality in inactive/standby state/mode before it is allowed to be used along with the inference of the currently used functionality (R4-2412021)
  + Option 6: RAN4 to send an LS to RAN2 and RAN1 WGs on a need to consider conformance testability of AI/ML enabled functionalities when designing such LCM operations as monitoring/activation/update/fallback for the devices already deployed in the field. (R4-2412021)
  + Option 7: Others
* Recommended WF

Options should be discussed separately to see if any agreements can be reached

### Sub-topic 1-2

*Channel modeling/ testing data*

The issue of whether synthetic channels would be adequate for performance assessment/testing was brought up. Some companies also brought up the need to have field data or mechanisms to ensure that synthetic channels also work well in the field.

**Issue 1-2: Channel modeling**

* Proposals
  + Option 1: Use synthetic channels as the default assumption for both training and inference phases.
  + Option 2: RAN4 must conduct an analysis for each use case to determine the reliability of using synthetic channels for test data in evaluating models trained on real data.
  + Option 3: RAN4 should discuss whether/how field data can be used for testing
  + Option 4: Other
* Recommended WF

Multiple options can be discussed

### Sub-topic 1-3

*Testing goals*

The testing goals have been discussed in previous meetings based on the options previously agreed and captured in the TR. The main discussion point is whether the requirements and tests should be for a given model or just for a functionality/feature.

**Issue 1-3: Testing goals**

* Proposals
  + Option 1: Requirements should be defined on a functionality level and DUTs can select any model to pass the test. The testing goal definition should be unified as follows:
    - The testing goal is to verify whether the minimum performance of AI/ML functionality/feature can be achieved in different test configurations/parameters.
  + Option 2: Deprioritize Option 1 and focus on Option 2
    - RAN4 can define multiple independent test case with different scenario and configurations/conditions depending on UE capability
  + Option 3:
* Recommended WF
  + To be discussed

### Sub-topic 1-4

*AI/ML performance*

Multiple companies brought up the issue that AI/ML performance should be better than legacy (in case there are legacy performance requirements for the same functionality/feature)

**Issue 1-4: AI/ML model complexity**

* Proposals
  + Option 1: AI/ML based performance should exceed legacy non-AI/ML requirements
  + Option 2:
  + Option 3:
* Recommended WF

To be discussed

### Sub-topic 1-5

*Generalization*

Ensuring generalization of the models used by the devices is very important for robustness.

**Issue 1-5: Generalization**

* Proposals
  + Option 1: If non-static scenarios/configurations are supported for certain use cases, they can be included as part of generalization tests.
  + Option 2: take the requirements for inference as the minimum level performance for generalization.
  + Option 3: For generalization test, performance for other scenarios and/or configurations can be ensured by RAN4 legacy non-AI test
  + Option 4: Others
* Recommended WF

To be discussed

### Sub-topic 1-6

*CR in R4-2412023*

**Issue 1-6: CR in R4-2412023**

* Proposals
  + Option 1: Proposed changes are summarized below

|  |  |
| --- | --- |
| ***Reason for change:*** | Introduced definitions are not applicable to conformace testing. |
|  |  |
| ***Summary of change:*** | Added a note to the definitions of the following terms:   * AI/ML model testing * AI/ML model validation |
|  |  |
| ***Consequences if not approved:*** | The definitions of the terms can be considred generic and one can confuse those with conformace testing practices. |

* + Option 2: CR is not agreeable
* Recommended WF

To be discussed. If CR should be revised, concrete proposals should be presented

### Sub-topic 1-7

*LCM handling*

**Issue 1-7: LCM handling**

* Proposals
  + Option 1: Postpone any LCM discussion until more RAN2/1 agreements are made.
    - agreements on possible requirements were made in the last meeting, all of them to be discussed later
  + Option 2: RAN4 to study a maximum delay requirement for switching between AI/ML and non-AI/ML operation modes, at least in the fallback direction (AI/ML →non-AI/ML).
  + Option 3: Other options
* Recommended WF

To be discussed

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2411177**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411177.zip) | Korea Testing Laboratory | **Observation 1: It costs a lot to consider all possible combinations between Set A and Set B for the conformance testing.**  ***Proposal 1: For the conformance testing, study and specify the reduced or a single combination(s) between Set A and Set B.***  **Observation 2: In the evaluation, UE rotation is modelled for BM-Case2 with a rotation speed in all three rotational axes, with the rotational direction chosen uniformly at random among the three axes.**  ***Proposal 2: The following issues should be considered***   * ***Study on the testing BM-Case2 whether or not the inference algorithm is independent of the rotational direction of UE.***    + ***Study the test setup on the existing FR2 OTA setup as baseline.***   + ***Otherwise, study a test environment considering the rotational direction of UE***   **Observation 3: Two kinds of AI/ML model identification shall be considered for test setup.**  **Observation 4: The field deployment condition includes the LCM function.**  ***Proposal 3: RAN4 shall consider whether or not the LCM performance monitoring function shall be involved since the field deployment condition reflects the LCM.***  **Observation 5: Required efforts are made to configure models to reflect real-field environments for AI/ML use cases, however, it's important to note that in the conformance testing, RAN4 sets minimum requirements for each test case.**  **Observation 6: While data augmentation is a powerful tool for mitigating overfitting, it may not completely prevent the issue in all cases when solely relied upon.**  ***Proposal 4: It is necessary to validate whether test configurations, especially those considering data augmentation, can effectively prevent overfitting. If there is uncertainty, other approaches should be taken into account.***  **Observation 7: The measurement error degrades the performance of the legacy and also AI/ML beam management.**  ***Proposal 8: Study the possibility of enhancing the area to have a higher level of accuracy than the current FR2’s L1-RSRP accuracy.*** |
| [**R4-2411254**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411254.zip) | Keysight Technologies UK Ltd | *Observation 1: The AI/ML beam management objective seems primarily targeted for FR2.*  *Observation 2: Only the FR2 MIMO OTA test system can properly and accurately implement CDL channel model.*  *Observation 3: All existing FR2 test systems could properly and accurately implement a TDL channel model.*  *Observation 4: The Enhanced IFF and FR2 MIMO OTA test systems can realize multiple TRPs (with TDL channel models at most) simultaneously.*  *Observation 5: Only the IFF and Enhanced IFF truly present FF conditions.*  *Observation 6: The default FR2 MIMO OTA test system cannot present two vastly dislocated TRPs with CDL channel models simultaneously.*  *Observation 7: Depending on the angular separation of the TRPs, system upgrades of existing FR2 MIMO OTA systems could increase cost&complexity and potentially the chamber size.*  *Observation 8: Multiple TRPs can be presented to the UE sequentially by rotating the UE but the positioning time in excess of 1s might be too slow.*  *Observation 9: Multiple TRPs can be presented to the UE sequentially without UE positioner movements with either the Enhanced IFF or FR2 MIMO OTA test system with insignificant delays.*  *Observation 10: It is advisable to limit the maximum number of spatially separated TRPs to 2 if more than 1 TRP is required*  *Observation 11: CDL models are realistic channel models for beam management test cases while TDL models which cannot reflect any realism for the beamforming effects at all.*  *Observation 12: Multiple strong gNB beam candidates for beam selection in multipath propagation environments exist.*  *Observation 13: Each gNB beam candidates can result in different arrival PAS for the UE reception, potentially impacting also UE beam selection.*  *Observation 14: Beamforming performed at both link ends (gNB and UE) filters out weak multipath clusters of the channel model, hence applying gNB beamforming simplifies the resulting channel model seen by the UE to a reasonable number of strong clusters.*  *Observation 15: gNB and UE beam selection impacts resulting cluster-wise fading characteristics.*  *Observation 16: To properly assess AI/ML algorithms, the complexities of beam dependent fading characteristics and realistic channel models need to be taken into account.*  *Observation 17: The performance of beam management operations depends strongly on the spatial characteristics of the multipath propagation environment which can be accurately assessed in the FR2 MPAC OTA system.*  *Observation 18: A spatial channel model is required as a basis of test condition definition to determine L1-RSRP relations between the beams of different sets.*  *Observation 19: The reference L1-RSRP relations between the beams can be determined using the channel modelling methodology defined in TR 38.901 with parameters and details based on TR 38.827/TS 38.151 after the Tx-beams and Rx-beams are defined*  *Observation 20: From a purely channel model perspective and the realization in state-of-the-art channel emulators, implementation of CDL models is not more complex than TDL models*  *Observation 21: The RX PAS of 4 strongest TX beams can be emulated reliably and accurately with 8 probes*  *Observation 22: A relatively simple and straightforward extension of an FR2 MIMO OTA system to 8 probes is suitable to emulate 4 Tx-beam RX-PAS profiles.*  *Observation 23: To emulate set B Tx beams beyond 4 [up to 16], power offsets can be applied to the Tx-beams with the closest match of the 4 available RX-PAS profiles.*  *Observation 24: Various beam management performance testing can readily be supported with a commercially available test system, i.e., complexity and time-to-market concerns should not be an issue.*  *Observation 25: Performing conformance testing with TDL models that are far less realistic and that the ML model likely has not been trained on due to the lack of realism is likely to provide inconsistencies in test results.*  **Proposal 1: It is proposed to limit the testability discussions of AI/ML Beam Management test cases to FR2 only.**  **Proposal 2: Take the FR2 OTA test systems and applicabilities in Table 1 into account for the testability discussions of AI/ML Beam Management.**  Table 1: Applicability of existing OTA test systems to testability criteria for AI/ML beam management   |  |  |  |  | | --- | --- | --- | --- | | **Applicability Criteria** | **IFF**  **A diagram of a motor coordinates  Description automatically generated** | **Enhanced IFF**  **A diagram of a machine  Description automatically generated** | **FR2 MIMO OTA**  **A grey round object with a green arrow and several blue lights  Description automatically generated with medium confidence** | | **CDL Channel Models** | No | No | Yes | | **TDL Channel Models** | Yes | Yes | Yes | | **Multiple TRPs simultaneously** | No | Yes | Yes | | **True FF Conditions** | Yes | Yes | Not necessarily | | **Multiple TRPs with CDL channel models presented simultaneously** | No | No | Requires upgrade whose cost & complexity might be not negligible | | **Multiple TRPs with TDL channel models presented simultaneously or sequentially at *t*0 and *t*0 + D*T* (without positioner movement)** | No | Yes | Yes | | **Multiple TRPs with CDL channel models presented sequentially at *t*0 and *t*0 + D*T* (using positioner movement)** | No | No | Yes, but D*T≥*1s might be excessive | | **Multiple TRPs with at most TDL channel models presented sequentially at *t*0 and *t*0 + D*T* (using positioner movement)** | Yes, but D*T≥*1s might be too slow | Yes, but D*T≥*1s might be too slow | Yes, but D*T≥*1s might be too slow |   **Proposal 3: Infra vendors to clarify the details around gNB beam candidates, including the minimum number of beams, beam steering capabilities/codebook, latency, that should be presented to the UE.**  **Proposal 4: Consider the FR2 MIMO MPAC OTA test system as the baseline for AI/ML Beam management - DL Tx beam prediction use case.**  **Proposal 5: Consider an 8-probe OTA system with 2 additional probes when compared to the FR2 MIMO OTA system as the starting point for AI/ML Beam management testing for 1 TRP and/or 2 TRPs with close angular proximity.** |
| [**R4-2411259**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411259.zip) | CAICT | **Proposal 1: For RSRP prediction accuracy, suggest to prioritize option 3 and 1, and potential test burden with each option in future discussion.**  **Proposal 2: RAN4 to study the possibility of reflecting the effect of diverse Tx beams on receiver side through identifying a set of propagation conditions in channel model.**  **Proposal 3: RAN4 to discuss whether to introduce the testing feasibility and possible methodology for AI/ML beam management into Rel-19 study item on NR FR2 OTA testing enhancement.**  **Proposal 4: if testing datasets is adopted, suggest to specify basic structural parameters to unify input format of inference test.**  **Proposal 5: Suggest to discuss whether the overfitting problem can be relieved via generalization test.**  **Proposal 6: Hold on the discussion on consistency until RAN1 has sufficient progress on this.** |
| [**R4-2411279**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411279.zip) | Anritsu Corporation | ***Observation 1: Due to the total number of clusters, it is not practical to apply the existing CDL channel model directly for the beam management verification test.***  ***Observation 2: In a case of NLOS CDL channel model, there are many clusters (angle of arrivals) from 5 to 10 cluster groups (requiring from 5 to 10 TE probes) even after reducing and grouping them.***  ***Observation 3: In a case of LOS CDL channel model, clusters can be reduced down to two cluster groups (requiring two TE probes).***  ***Observation 4: The number of clusters must be minimized as much as possible to avoid an overly complex test system with its associated drawbacks.***  ***Observation 5: It must be noted that two polarizations are needed for each test probe per cluster.***  ***Observation 6: Clusters are arranged in a wide range of azimuth angles from -180 to 180 degrees, which may increase the difficulty for arranging the actual probes in an FR2 OTA chamber.***  ***Observation 7: A footprint of the OTA chamber may become huge if we need to arrange probes over the whole x-y plane and its proximity while fulfilling the far-field criteria.***  ***Observation 8: Care must be taken not to place probes facing each other to avoid interference and degradation of QoQZ.***  ***Observation 9: The total number of probes may increase to support all the CDL channel models.***  ***Observation 10: It is expected that a procedure to change probe locations cannot be allowed due to concern of output power calibration errors in FR2.***  ***Observation 11: To minimize the number of probes, further study and clarification are necessary with the following main questions to be answered.***   * ***How many and which CDL channel models need to be supported for the beam management test?*** * ***How many clusters are necessary in each channel model?*** * ***How much antenna offset is necessary to avoid the interference and degradation of QoQZ?*** * ***What conditions can we apply to achieve minimum number of clusters per channel model?***   ***Observation 12: Grouping by assigning antenna probes to different clusters having AOA offset below the threshold set as +/-15 degrees, even though preservation of original delays and Doppler frequencies, AoAs reproduction in conformance testing will be impacted particularly if many CDL models need to be supported each having many clusters and different AoAs. A restricted number of CDL models developed for conformance testing while realistic would be ideal.***  ***Proposal 1: RAN4 should decide what are the CDL channel models necessary for testing and limit its number as much as possible. This will help determining what is possible from a test point of view.***  ***Proposal 2: RAN4 should model new CDL channel model parameter sets taking into account of the conformance testing, and group/ reduce number of clusters and angle of arrivals.*** |
| [**R4-2411292**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411292.zip) | OPPO | **Proposal 1: To refine the definition of *RSRP accuracy*, following options should be further clarified,**   * **Option 1: the RSRP difference between the ideal(or measured) L1-RSRP of Top-1 predicted beam and the ideal(or measured) L1-RSRP of the Top-1 ideal beam** * **Option 2: the RSRP difference between the predicted L1-RSRP of Top-1 predicted beam and the ideal(or measured) L1-RSRP of the Top-1 ideal beam** * **Option 3: the RSRP difference between the predicted L1-RSRP of Top-1 predicted beam and the ideal(or measured) L1-RSRP of the Top-1 predicted beam**   **Proposal 2: Regarding the model output tests, the selection of metrics depends on the progress in RAN1,**   * **if *Beam information on predicted Top K beam(s) among a set of beams* is the only reported content, beam prediction accuracy may be the only available KPI accordingly (also depends on the refined definition of *RSRP accuracy* in RAN4),** * **if both *Beam information on predicted Top K beam(s) among a set of beams* and *RSRP of predicted Top K beam(s) among a set of beams* are reported by the UE, both beam prediction accuracy and RSRP accuracy could be utilized as RAN4 test KPIs.**   **Proposal 3: UE can monitor the performance of AI/ML based CSI model/functionality through RSRP accuracy or Beam prediction accuracy, depends on the progress in RAN1.**  **Proposal 4: Stability of the performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes.**  **Proposal 5: For BM testability, RAN4 need to:**   1. **Clarify the limitations regarding the FR2 beams or probes that TE vendors could support** 2. **To determine the test setup for BM, a potential approach could be:**  * **Assuming that TE supports X probes in a chamber** * **By utilizing X probes, X different angles of arrival (AOA) can be achieved and can be mapped to a TX beam transmission with X paths(clusters)** * **Different TX beams in BM set A or set B can be represented by setting different pathloss for the different AOA paths, to mimic the beam sweeping procedure** * **FFS the number of probes that could be supported by TE vendors** * **FFS the number of AoAs that RAN4 needs to represent a beam** * **FFS how to map the X AoAs and corresponding X pathloss to different beams in BM set A/B**     **Proposal 6:**   * **Maximum number of set B Tx beams that test system should be able to emulate: 8** * **Maximum number of set A Tx beams that test system should be able to emulate: 64**   **Proposal 7: Further consider how to ensure that the BM model constructed on the DUT side can match(or approximate match) and be utilized in the testing environment on the TE side.** |
| [**R4-2411341**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411341.zip) | CATT | **Proposal 1: RAN4 not to down select any metric/KPI for BM use case until RAN1 have progress on performance monitoring metric.**  **Proposal 2: Discussions on Option 2 and Option 3 can be postponed unless feasible solutions are provided.**  **Proposal 3: RAN4 continue to discuss Option 1/4/5 and down-selection can be made when RAN1 have agreements on performance monitoring metric.**  **Proposal 4: Augmenting training data to avoid over-fitting to the test environment is left to UE implementation.**  **Proposal 5: RAN4 not to define dataset for testing unless the motivation is clarified.**  **Observation 1: RAN1 studied the impact of measurement error on the inference accuracy in SI phase.**  **Proposal 6: RAN4 not to repeat the RAN1 study on the impact of measurement error.**  **Proposal 7: Two solutions can be considered for defining performance requirements:**  **Alt 1: A margin is added to the accuracy requirements which are derived based on datasets without error or to the existing accuracy requirements if it is agreed to reuse the existing ones. The margin can refer to the RAN1 evaluation results.**  **Alt 2: The accuracy requirements are newly defined based on the datasets with error.**  **Proposal 8: RAN4 not to discuss to introduce a different reporting scheme.** |
| [**R4-2411409**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411409.zip) | Apple | **Observation 1: The testing environment for Beam Management (BM) case 1 and BM case 2 should replicate conditions that sufficiently capture the correlations of the transmit (Tx) and receive (Rx) beam patterns across the entire spectrum of propagation conditions in both spatial (angles of arrival and departure - AoAs and AoDs) and temporal domains.**  **Observation 2: The following conditions will introduce randomness and variations in propagation conditions across both time and spatial domains for the computation of L1-RSRP**   1. **Different AoDs with respect to the Tx antenna array** 2. **Different AoAs with respect to the Rx antenna array** 3. **Different superpositions of {AoA,AoD} pairs** 4. **Fading/Variation in time domain ( different {AoA,AoD} pairs per resolvable delay bin path)** 5. **UE movement (including rotation)**   **Observation 3: During real-world deployment, UE will encounter random radio propagations characterized by variations in both spatial and temporal domains (fading). Testing UE under similar conditions is important to reflect the realities of deployment accurately.**  **Observation 4: The following questions need to be answered to evaluate the feasibility of the FR2 OTA-based test procedure:**   1. **How can we generate multiple beams from the Set-B and Set-A Tx beams given the limitation of two AoAs?** 2. **What assumption is made regarding Rx beam sweeping? Does the UE utilize a fixed Rx beam, or does it sweep to find the optimal Rx beam?** 3. **How can we simultaneously emulate different AoAs (Rx beam) and AoD (Tx beams)?** 4. **How can we achieve dynamic variation in the AoD domain (Tx beam sweeping) for BM case 2 prediction?**   **Observation 5: From the network operator's perspective, a number of optimal beams need to be known in order to optimize load balancing and make trade-offs in performance and complexity**  **Observation 6: Ensuring that RSRP accuracy implies beam prediction accuracy is challenging and depends on absolute RSRP accuracy and the RSRP difference between the best beams**  **Observation 7: When using neural networks for regression, we rely on the Normalized Mean Squared Error (NMSE) criterion for training. It's crucial to ensure that the accuracy of weaker beams is not compromised by the accuracy of the strongest beams**  **Observation 8: The source of training data for beam management will play a crucial role in AI/ML BM performance and in the generalization performance in real deployment**  **Observation 9:** **For training data based on real measurements, the** **quality of training data depends on RF impairments, and other noise sources. There is tradeoff between training data quality and generalization performance. With training data collection from the field (across varying SNR conditions), both set-B and set-A beams will be affected ny impairments**  **Observation 10:** **To guarantee that the UE operates within acceptable margins, it's essential to subject it to various radio conditions and additional conditions for testing and generalization validation in RAN4**  **Observation 11:** **For BM use case the identified scenarios and configurations can be initially understood as those reported by UE through capability signaling as part of functionality identification.**  **Observation 12: The additional conditions for the AI/ML model training (which do not constitute part of UE capability) for the AI/ML-enabled feature/FG can serve as the different scenarios/configuration for defining generalization**  **Observation 13: Achieving consistency between training and inference by model monitoring could result in delays and increased complexity in model management for BM use case**  **Observation 14: If multiple models with varying generalization capabilities and requirements for network-side additional conditions are trained by different UE vendors, it would necessitate substantial standardization efforts for BM use case**  **Observation 15: Current proposals on assistance information for additional conditions and Model Identification only serve the purpose of selecting the appropriate AI/ML model. However, this approach may not be scalable due to considerations of UE implementation complexities and granualtity of conditions/additional conditions. Complexity can increase substantially, especially if condition granularity is fine.**  ***Framework for Defining Requirements in Tests for AI/ML***  **Proposal 1**: **RAN4 should study the specification of reference AI/ML models, training procedure, and training data for defining performance requirements for BM use case while considering limitations on model complexity. We provide a reference table for initiating the discussion.**  ***Beam Prediction Testability Discussion***  **Proposal 2:** **RAN4 should investigate the feasibility of the current FR2 OTA-based test procedure to capture random fading in both spatial and temporal domains, similar to CDL, Uma, etc. Additionally, considering the incorporation of UE rotation can help model randomness from the perspective of the Angle of Arrival (AoA)**  **Proposal 3: If the current FR2 OTA-based test procedure proves to be not feasible, consider using CDL and Uma channels for testing in the BM AI/ML use case upon TE approval for feasibility**  **Proposal 4: For testing the KPIs for BM-Case 1 spatial prediction we propose the test setup and framework described below upon TE vendor approval for the feasibility of this approach.**  **A diagram of a test  Description automatically generated with medium confidence**  **A diagram of a diagram of a diagram  Description automatically generated with medium confidence**  **Proposal 5: For testing the KPI for BM-Case 2 temporal prediction we propose the test set up and framework described below upon TE vendor approval for the feasibility of this approach.**  A diagram of a graph  Description automatically generated  ***KPIs/Test Metrics for BM Use case***  **Proposal 6: We propose that RSRP accuracy be defined as the difference between the predicted RSRP and the measured (genie) RSRP associated with the same Tx beam.** **Accuracy requirements should apply to all predicted beams that satisfy the predefined side conditions, including SNR.**  **Proposal 7: When considering the necessity of additionally testing for beam prediction accuracy, we propose to study the additional information and significance that this test will provide, especially in light of our definition of RSRP accuracy.**  ***Datasets for Training/Testing***  **Proposal 8:** **Regarding the source of training data, RAN4 should agree that the training dataset is for testing purposes only. In real-world scenarios, a different training dataset could be obtained to train a vendor-specific model, unless vendors are willing to share their codebooks. In such a case, we could define a mixed training dataset from multiple vendors. Further feasibility studies (FFS) are needed to determine if this approach is viable.**  **Proposal 9: For the source of the training data for BM, RAN4 should define the channel conditions and configurations to be used in the testing environment. Additionally, UE-generated training data should take into account each unique hardware implementation (Option 3, combination of Option 1 and 2).**    **Fig. 7: Option 3 for generating the training dataset for beam management**  ***Measurement Acuracy for BM***  **Proposal 10: To investigate performance degradation resulting from measurement accuracy, we propose specifying an impairment model along with its associated parameters. This ensures that companies provide comparable simulation results, which will guide the decision on changing RSRP accuracy requirements.**   |  |  |  | | --- | --- | --- | | **Category** | **Parameter** | **Description/Examples** | | Training Data for BM | Training datasets | System level channel model (CDL,Uma, etc) number of set -B/ set-A beams, specific TX and RX codebooks.  Antena port layout  Best Rx beam selection  SSB Wide Tx beams vs CSI-RS Narrow beams  Tx Codebook  Tx antenn array dimensions  Antenna Element Spacings | | Impairment Model to be added to set B and set A measurmenets | Impairment types affecting set B measurements and set A measurements, quantization, SNR levels, RF impairmenets, other side conditions etc |   **Fig. 8: Proposed parameters to generate training data with impairment model for BM**  **Proposal 11: If the outcome of simulating the performance with the impairment model results in degraded performance, then we can consider the following options:**   1. **Adjusting RSRP Requirements and Enhance Measurement Accuracy: Reevaluate and potentially modify the current RSRP accuracy requirements to better align with the observed performance under the impairment model.** 2. **Adopting Compensation Mechanisms: Make the AI/ML adaptable to changing SNR conditions, for example to maintain performance the number of set B beams can increase, also the number of K in top-K could increase to compensate for degraded performance. (change model, address scalability of AI/ML model)** 3. **Change the side conditions, (like SNR, etc) under which the AI/ML operation can be supported. Below a threshold SNR point, legacy procedures should be employed.**   ***Generalization issues for BM***  **Proposal 12: In the RAN4 core requirement, it is mandated that the consistency or association between Set B beams and Set A beams during both training and inference must be guaranteed. This serves as additional assistance information from the network side for testing UE-sided models during inference and monitoring**  **Proposal 13: RAN4 should define identified scenarios/configurations associated with the UE capability report of an AI/ML-enabled Feature FG. For defining generalization tests, the additional conditions can serve as the other identified scenarios/configurations for the BM use case**  **Proposal 14: RAN4 should investigate the feasibility of providing assistance information for the additional conditions to aid generalization and consistency across training and testing when defining requirements. Other additional conditions that are not part of UE capability can be used to define generalization tests**  **Proposal 15:** **For additional conditions that cannot be shared due to proprietary concerns, RAN4 can explore the feasibility of using a virtual ID to indicate the specific conditions under which a model was trained. This approach would assist in the proper selection of UE models to support generalization. Additionally, RAN4 should identify which additional conditions should be exclusively reserved for generalization tests.**  ***Consistency between Training and Inference***  **Proposal 16: RAN4 to consider option 1 as a baseline for achieving consistency between training and inference for both set A and set B for beam spatial and temporal prediction**  **Proposal 17: RAN4 to consider option 1 for achieving consistency between training and inference for only static conditions.**  **Proposal 18:  In order to ease the burden for testing models with different NW additional conditions, it would beneficial to train the UE-side model with mixed dataset from various gNB settings, thus reducing the number of AI/ML models (selected by NW-side additional conditions) required to guarantee generalization and maintain the system performance for BM use case**  **Proposal 19: For achieving consistency between training and inference, if the model drifts due to misalignment of network-side additional conditions, the alignment could be achieved through an ID assigned by the network during training data collection. This ID indicates the association of the training data with the additional conditions implied to generate those data**  **Proposal 20: For achieving consistency between training and inference, if the UE lacks the ID but the network possesses the model ID, then the network can transfer the model to the UE, and the UE can update its list of models**  **Proposal 21: For achieving consistency between training and inference, if the UE has a model that partially supports the network-sided additional conditions, then the network can trigger data collection for fine-tuning the UE-sided model. Subsequently UE updates its list of model IDs.**  **Proposal 22: The UE updates its model list with a new model derived from either fine-tuning, model transfer from the NW, or monitoring to ensure the consistency of additional conditions. Then, the UE assigns an ID to the new model that supports the NW's additional conditions and shares this information with the NW. If some of the new conditions/configurations are standardized, the UE updates its capability report accordingly.**  **Proposal 23: If there is no ID available to associate training data with additional conditions, and monitoring procedure fails to guarantee the consistency of the model with the additional conditions then UE should fallback to legacy mode.**  **Proposal 24: Investigate the feasibility of enhancing the generalizability of the AI/ML model and reducing the number of AI/ML models and the testing burden for the beam management case by supplementing the core AI/ML input signals with both network (NW) and UE auxiliary information signals integral to its inference engine**    **Proposal 25: Investigate the feasibility of training the models with a mixed dataset associated with both network (NW) and UE auxiliary information signals to further enhance the generalizability of the AI/ML model for the beam management case and reduce the number of generalization tests.** |
| [**R4-2411587**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411587.zip) | Rohde & Schwarz | Observation 1: The current FR2 NR MIMO OTA system layout is limited to 6 probes, within a 55º by 10º sector, to simulate the resulting AoA/ZoA after applying the BS filtering (i.e. strongest beam towards the strongest cluster) for one, from two possible simplified CDL channel modes (i.e. CDL-A InO and CDL-C UMi).  Observation 2: The near field distribution in the test zone, simulating a complex multipath far field environment for the FR2 NR MIMO OTA system, is the result of combining the simultaneous signals transmitted by all probes.  Observation 3: The current FR2 NR MIMO OTA system defined in TR 38.827 cannot be used for testing AI/ML based beam management scenarios.  Observation 4: Upscaling the FR2 NR MIMO OTA methodology (based on 3D MPAC chamber) to emulate CDL-C UMi requires a test system with >100 probes and >200 RF channels which is unfeasible.  Observation 5: For feasible implementation of an OTA test setup for AI/ML based beam management, further simplifications are inevitable.  Observation 6: Key requirement for AI/ML based beam management testing is not the CDL channel as such, but a time varying and accurate DL power, resulting as combination of signals from various impinging angles, which is suitable for accurate and reliable (L1‑)RSRP measurements.  Observation 7: Multi-AoA systems-based on IFF (e.g. Enhanced IFF) provide better DL power accuracy and lower MU compared to DFF based systems and thus, represent the best candidate to address the power accuracy based requirements for AI/ML based beam management testing.  Proposal 1: RAN4 to consider a sparse probe layout as baseline for AI/ML based beam management testing and discuss the testability requirements for:   * Set of CDL models to be considered. * Minimum number of clusters to be emulated for the CDL model(s), that would not bias the results (e.g. by power truncation). * Neglection of the angular spread of the clusters in the spatial domain (i.e. single AoA/ZoA per cluster). * Limitation of the angular range (azimuth and elevation).   Proposal 2: RAN4 to consider multi-AoA IFF based system as baseline for AI/ML based beam management testing. |
| [**R4-2411626**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411626.zip) | Xiaomi | **Observation 1: From TR, predicted best TX beam index and/or L1-RSRP can be based on full RX beam sweeping or specific RX beam.**  **Observation 2: Whether specific RX beams or full RX beam sweeping is assumed may lead to different best predicted TX beam index and predicted L1-RSRP value.**  **Observation 3: In legacy non-AI based L1-RSRP measurement, when RX beam sweeping is applied, it’s up to UE implementation to choose RX beam angle.**  **Observation 4: If different RX beam is assumed for prediction and measurement, the L1-RSRP and/or best TX beam index may be different in two steps, which will lead to wrong performance evaluation.**  **Proposal 1: RAN4 to discuss how to keep the RX beam Consistency between prediction and measurement for set A to correctly evaluate prediction performance.**  **Proposal 2:** **Suggest to use RSRP delta CDF-based simulation to determine the appropriate SNR level for deriving ground truth.**  **Proposal 3: For BM case-1, channel doppler can set to 0 or a small value to guarantee that there is neglectable L1-RSRP variation.**  **Proposal 4: for BM case-2, channel doppler will depend on UE speed and UE trajectory. RSRP variation impact can be considered in RSRP accuracy requirement.**  **Proposal 5: It’s more challenging to obtain accurate RSRP measurement in multiple-AOA test setup. it’s FFS whether extra margin needs to considered in RSRP accuracy requirement.**  **Observation 5: UE will report predicted RSRP/beam at T1 and measured best beam as ground truth at T2 to TE, it’s easy for UE to pass the test by cheating.**  **Proposal 6: For test metric option 1, refine the definition:**   * **Option 1: Absolute RSRP prediction accuracy is the RSRP difference between the predicted L1-RSRP of Top-1 predicted beam and measured L1-RSRP of the same beam.**    + **FFS to define relative accuracy requirement.**   **Proposal 7:** **For test metric option 3, refine the definition:**   * **Option 3: successful rate for the correct beam prediction which is considered as measured L1-RSRP of top-1 predicted beams is larger than the measured L1-RSRP of the Top-1 measured beam – x dB.**   **Observation 6: From RAN1 agreement, both predicted beam index and predicted RSRP can be included in L1 report at least for BM case-1, similar as legacy.**  **Observation 7: For option 1, RSRP prediction accuracy focus about the RSRP difference for the same beam.**  **Observation 8: Both option 2 and 3 can be classified into beam prediction accuracy. Option 2 refers to beam index difference while option 3 refers to RSRP difference between predicted best beam and ideal best beam.**  **Observation 9: RSRP prediction accuracy requirement can’t replace beam prediction accuracy.**  **Proposal 8: RAN4 needs to at least define beam prediction accuracy requirement and discuss which test metric to be chosen, e.g. based on beam index difference or RSRP difference.**  **Observation 10: When best predicted TX beam is in set B, UE knows the best RX beam. When best predicted TX beam is in set A but not in set B, it’s still possible for UE to know the RX beam corresponding to the best TX beam.**  **Proposal 9: RAN4 to discuss known TCI state condition for AI based beam indication.** |
| [**R4-2411635**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411635.zip) | Samsung | **Proposal 1: RAN4 shall firstly agree on the expected RAN4 requirement impacts for AI-BM, by separating the discussion by (1) NW and UE-sided model; (2) different sub-features introduced in RAN1/2, including: Data collection, Inference, Perf. monitoring, LCM and additional assistance information (for UE-sided model only).**  **Proposal 2: The following table can be used as the template for collecting views from companies.**   |  |  |  | | --- | --- | --- | |  | New Sub-Feature introduced in RAN1/2 | Whether/How RAN4 requirement impact is expected? (e.g., clause # in TS 38.133 impacted/new requirement needed/no impact) | | NW-sided model | Data collection |  | | Inference |  | | Perf. monitoring |  | | LCM |  | | UE- sided model | Data collection |  | | Additional assistance information |  | | Inference |  | | Perf. monitoring |  | | LCM |  |   *RAN4 Core Requirements for Supporting NW-sided Model*  **Proposal 3: RAN4 shall introduce the necessary L1-RSRP reporting requirement on supporting data collection for NW-side AI/ML model inference/training (for both BM-Case1 & 2), by considering:**   1. **L1 measurement/reporting for inference: e.g., beam reporting for more than 4 beam in L1 signaling, and other new reporting format (including potential granularity step);** 2. **Potential enhancement on MDT-based measurement/reporting for training.**   **Proposal 4: For tightening measurement accuracy requirement for AI-BM NW-sided model:**   1. **RAN4 shall only focus on the feasibility part of tightened measurement accuracy;** 2. **Without any particular request, RAN4 shall reuse the accuracy requirement from Rel-15 L1-RSRP for both absolute and relative RSRP reporting.**   **Proposal 5: No RAN4 impact is expected for NW-side AI/ML model inference for BM-Case1 & 2, except data collection for model inference.**  **Proposal 6: FFS RAN4 impact from NW-side AI/ML model performance monitoring for BM-Case1 & 2 (depending on RAN1/2 input).**  **Proposal 7: No RAN4 impact is expected for NW-side AI/ML model LCM for BM-Case1 & 2.**  *RAN4 Core Requirements for Supporting UE-sided Model*  **Proposal 8: FFS RAN4 impact from data collection for UE-sided model training (depending on RAN1/2 input).**  **Proposal 9: For the testing of UE-sided model inference and monitoring (if any), the consistency of NW-side additional condition across training and inference shall be guaranteed as side condition.**  **Proposal 10: RAN4 requirement of UE-side AI/ML model inference needs to be specified.**  **Proposal 11: RAN4 shall specify RRM requirement for UE-side AI/ML model inference: to guarantee (1) accuracy of inference results (2) the delay of reporting inference results (containing the inference latency).**  **Proposal 12: For different types of performance monitoring for UE-sided model, the necessity of RAN4 requirement is provided as:**   1. **Type 1, Option 1 (NW-side performance monitoring): The necessity of RAN4 requirement on data collection for monitoring is not significant, because it is similar to data collection for other purposes.** 2. **Type 1, Option 2 (UE-assisted performance monitoring): RAN4 requirement on data collection for monitoring can be specified to test the accuracy of performance metrics calculated by UE.** 3. **Type 2 (UE-side performance monitoring): No RAN4 requirement is needed because no UE feedback will be performed.**   **Proposal 13: For UE-sided model LCM, the necessity of RAN4 requirement:**   1. **“Network decision, network-initiated” LCM: RAN4 requirement is needed.** 2. **FFS the detailed RAN4 requirement (depending on RAN1/2 input).**   *Summary of Proposals for AI-BM Core Requirement*  **Proposal 14: Based on corresponding proposals, the expected RAN4 requirement impact from AI-BM is summarized as:**   |  |  |  |  | | --- | --- | --- | --- | |  | New Sub-Feature introduced in RAN1/2 | Whether/How RAN4 requirement impact is expected? (e.g., clause # in TS 38.133 impacted/new requirement needed/no impact/FFS) | Related Proposal(s) | | NW-sided model | Data collection | * Clause 9.5 L1-RSRP measurements for reporting  (or similar new requirement for L1-RSRP measurement for reporting for AI-BM) * FFS new clause for MDT-based measurement/reporting for training * FFS Clause 10.1.20 for L1-RSRP accuracy requirements for FR2 | Proposal 3/4 | | Inference | * No RAN4 impact expected | Proposal 5 | | Perf. monitoring | * FFS new clause for performance monitoring | Proposal 6 | | LCM | * No RAN4 impact expected | Proposal 7 | | UE- sided model | Data collection | * FFS new clause for data collection for UE-sided model training | Proposal 8 | | Additional assistance information | * Added as new side condition: consistency/association shall be guaranteed for inference and perf. monitoring requirement testing | Proposal 9 | | Inference | * New clause for UE-side AI/ML model inference | Proposal 10/11 | | Perf. monitoring | * FFS new clause is needed for Type 1 - Option 1 * New clause for Type 1 - Option 2 * No RAN4 impact expected for Type 2 | Proposal 12 | | LCM | * New clause for RAN4 RRM requirement for “Network decision, network-initiated” LCM | Proposal 13 |   *Testability Issues for AI-BM*  **Proposal 15: For metrics for beam management requirements/tests, RAN4 adopt the following combined option of Option 2 and 3 for compromise:**   * **Option-New: The successful rate for the correct prediction which is considered as "the Top-1 predicted beam is one of the Top-K strongest beams, and the Top-1 predicted beam’s ground truth RSRP value is larger than the RSRP of the strongest beam – x dB”**   + **FFS the values of K and x**   **Proposal 16: The following steps can be followed to emulate the required channel model in OTA chamber for UE-sided AI-BM model inference requirement testing:**  **Step-1: Determine the required evaluation scenario;**  **Step-2: SLS for channel model for all TX beams;**  **Step-3: Test signal is mapped over probe(s) in OTA chamber.**  **Proposal 17: RAN4 shall firstly discuss the required evaluation scenario, which is used to perform SLS for generating the UE received signals from different beams, in order to obtain the dataset for AI model training/evaluation.**  **- The evaluation assumption from Table 6.3.1-1 in TR38.843 could be used as the baseline for RAN4 to perform SLS for generating the UE received signals from different beams.**  **Proposal 18: FFS the feasibility of DFF chamber and 3D MPAC in a separate OTA testability study item, by considering:**  **- The necessity/feasibility of evaluate RX beam management in OTA chamber;**  **- The feasibility to generate the required channel model (for certain TX beambook) with limited number of probes.** |
| [**R4-2411706**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411706.zip) | MediaTek Inc. | **Proposal 1**: CDL channel or channel models in 38.901 shall be used in AI/ML BM test.  **Proposal 2**: The number of Tx beams that test system should be able to emulate is equal to number of set A Tx beams, that is [64-128].  **Observation 1**: None of the existing OTA test systems are able to emulate [64-128] Tx beams with CDL channel. It is costly and complex to use multiple probes to emulate multiple AoDs and AoAs of [64-128] Tx beams with CDL channel.  **Observation 2**: It is difficult to ensure reference RSRP is measured under similar channel conditions as in prediction phase.  **Proposal 3**: How to avoid UE cheating in the test, such as reports fake values according to the previous predicted results, shall be further discussed to use the reported RSRP measurement as ground truth.  **Proposal 4**: In AI/ML BM test, consider the basic model which uses L1 measurement results of Set B as input. Beam ID information may be used as input too.  **Proposal 5**: In AI/ML BM test, emulate the randomness and variations in propagation conditions across both time and spatial domains through adjusting Tx power at TE on each beam at different time.  **Proposal 6**: Only to test AI/ML BM with fixed Rx beamforming gain.  **Observation 3**: With single AoA, genie-aided relative strength of Tx beams is available at TE.  **Proposal 7**: Use existing IFF or enhance IFF test systems with single AoA for AI/ML BM test.  **Proposal 8**: What the reference RSRP is, measured or ideal RSRP, depends on the test setup finally adopted in RAN4.  **Observation 4**: Only DL Tx beam prediction is in scope and Rx beam prediction is not for both spatial-domain and temporal beam prediction.  **Proposal 9**: The TCI state QCL to an RS that is not in Set B is unknown if no L1 measurement is performed by UE within 1280ms before TCI state activation.  **Proposal 10**: The TCI state which is QCL to an RS in Set B is known:   * if the corresponding predicted beam is reported in 1280ms before the TCI state switch command and SNR of the RS is above -3dB for spatial-domain beam prediction. * if the last observation occasion is within 1280ms before the TCI state switch command and SNR of the RS is above -3dB for temporal beam prediction. |
| [**R4-2411978**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411978.zip) | CMCC | ***Proposal 1: for AI/ML based beam management, it is proposed that the*** ***absolute L1-RSRP accuracy are defined as the difference between the predicted RSRP and ground truth***   * ***For FR2, the ground truth is the approximate as the reported RSRP measurement under the certain SNR on the predicted Tx beam.*** * ***For FR1, the ground truth is the ideal measurement of RSRP on the predicted Tx beam.***   ***Proposal 2: for absolute RSRP accuracy, it is proposed that the RSRP difference is obtained from the same beam ID.***  ***Proposal 3: for Top-K case, it is proposed that RSRP accuracy is applied to all the K beams.***  ***Proposal 4: for the study of the impact of measurement accuracy on prediction accuracy, it is proposed to evaluate the performance impact due to measurement error with different value of K.***  ***Proposal 5: for the study of the impact of measurement accuracy on prediction accuracy, it is proposed to use legacy relative accuracy requirements for SSB based L1-RSRP measurement to model measurement error, i.e. ±3dB for FR1 and ±6.5dB for FR2.*** |
| [**R4-2412231**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412231.zip) | Ericsson | **Proposal 1: Metrics/KPIs for beam management requirements/tests separately in below options:**   * **In the case of beam information only in the content in the report of AI/ML model inference results**   + **The Metrics/KPIs for beam management requirements/tests must be based on the beam information.** * **In the case of beam information + RSRP in the content in the report of AI/ML model inference results**    + **The Metrics/KPIs for beam management requirements/tests must take both of the beam prediction accuracy and beam information into account, i.e., the UE shall fulfil beam prediction accuracy and beam information accuracy simultaneously.**   **Proposal 2: In the case of beam information + RSRP in the content in the report of AI/ML model inference results, the predicted RSRP accuracy of each beam in the Top-K predicted beams shall be individually verified with respect to the requirement to be defined.**  **Proposal 3: Relative RSRP accuracy, in same way of definition in legacy, shall be defined for Metrics/KPIs for beam management requirements/tests.**  **Proposal 4: For BM-Case 1, if a beam is configured for measurement in Set B, then the AI/ML model may produce the measured RSRP instead of the predicted RSRP. In this case, if beam information and RSRP both are in the content in the report of AI/ML model inference results, the Metrics/KPIs for beam management requirements/tests must take both of the measured RSRP accuracy and beam information into account, where measured RSRP accuracy shall follow the legacy requirements.**  **Proposal 5 RSRP prediction accuracy is defined as the difference between the predicted RSRP (of the predicted beam) and the ideal measurement of RSRP of the same Tx beam.**  **Proposal 6: The achievable maximal SNR level for ideal measurement of RSRP may need to be checked by test environment vendors.**  **Proposal 7: Adding ideal measurement of RSRP with high SNR condition into tests is an issue to be handled.**   * **It requests the UE to perform prediction/measurement with same channel condition (except for SNR condition) respectively for comparison of their RSRP.** * **Eventually, it may need a design between test environment and UE for a set of interactions and operations involving prediction, prediction report, increasing SNR, measurement, measurement report and comparison between measurement and prediction.**   **Proposal 8: Regarding the question on multiple-AoA test setup, channel condition, the ground truth will be changing or not, in our opinion, there is no specific treatment. They shall be the same as channel models for AI/ML beam management test.**  **Proposal 9: Take 3D Multi-Probe Anechoic Chamber (MPAC) for FR2 MIMO OTA testing as the start point of reference methodology for AI/ML tests.**  **Proposal 10: Take CDL channel realized by MPAC as the start point of reference channel for AI/ML tests.**  **Proposal 11: Regarding Channel generation in tests (for each test configuration, a test case may contain more than one test configuration):**   * **For BM-case1, channel change in time isn’t important. Therefore, generating channel #1 for training and generating channel #2 with same parameters for inference is enough.** * **For BM-case2, from test perspective, it needs some trackable change to mimic UE motion in test, which must rely on a set of consequent change of the channel parameters.**    + **At least, it includes generating a set of channels, from channel #1 to channel #m, for training and generating a set of channels, from channel #m+1 to channel #n, along with sequent and grade updates of a set of parameters is necessary.**      - **The number of m and n depend on the measurement window and prediction window of BM-case2.**     - **The parameters may comprise one or more than one of**        * **Departure angles (AOD, ZOD)**       * **Arrival angles (AOA, ZOA)**       * **Channel coefficient**       * **Others aren’t precluded**   **Proposal 12: RAN4 to define the maximum number of set B Tx beams = 32, maximum number of set A Tx beams = 256, as the starting point.**  **Proposal 13: At least two sets of number of set A and set B are provided. Below are two examples:**   * **For verifying UE computational power of the AI/ML model, one example is number of set A = 256, number of set B = 32.** * **For verifying UE performance (e.g., accuracy) of the AI/ML model, one example is number of set A = 256, number of set B = [8].** * **Different accuracy requirements may be applied for different sets of of number of set A and set B.**   **Proposal 14: Regarding AoAs, CDL channels may be the reference.**  **Proposal 15: Regarding UE rotation during the test, we can refer to definitions in MPAC, no UE rotation is needed.**  **Proposal 16: Regarding UE rotation/repositioning between different tests, we can refer to definitions in MPAC, no UE rotation/repositioning is needed.**  **Proposal 17: RAN4 to define the number of future time instances to predict, i.e., prediction window, for temporal beam prediction (BM-case 2).**  **Proposal 18: In order to avoid overfitting of UE models to the test environment only, the test (including data sets) shall be able to demonstrate the robustness of the AI/ML model.**  **Proposal 19: RAN4 to clarify whether training data set covers or doesn’t cover the hardware relevant error prior to the input of AI/ML model in UE.**  **Proposal 20: RAN4 shall study how to deal with hardware relevant error in training data. One option is taking the hardware relevant error into account the overall error.**  **Proposal 21: Consistency between training and inference shall cover below aspects:**   * **Same size of Set A of beams** * **Same size of Set B of beams** * **Same DL spatial TX-filter for each beam among of all Set A of beams** * **Same DL spatial TX-filter of for each beam among all Set B of beams** * **Same indexing/ordering of all Set A of beams** * **Same indexing/ordering of all Set B of beams** * **Same Set B pattern(s) (i.e., association between Set A and Set B)** * **Same QCL assumptions for resources configured, including Set B and Set A** * **Same deployment scenarios e.g., ISD, antenna height, down tilt and NLOS probability**   **Proposal 22: Consistency between training and inference shall consider the channel for test as follows:**   * **At least in BM-case2, the channel may vary between during training and inference, and UE’s position and rotation may vary as well.**   **Proposal 23: In case of AI/ML model with RSRP output, measurement error impact to prediction shall be verified, since the predicted RSRP is always worse than the measured RSRP since measurement error is introduced in training/inference phase of AI/ML model.**  **Proposal 24: Measurement error impact to prediction shall be verified (e.g. by simulation) to guarantee an acceptable prediction performance.**  **Proposal 25: If tightening measurement error under current side condition is difficult from UE implementation perspective, RAN4 can add another side condition for tightened measurement error for enabling AI-ML model.**  **Proposal 26: In case of AI/ML model with label (classifier network) output, measurement error may not have direct impact to the prediction, if the reporting of the UE only relies on the output of AI/ML model without explicit or implicit assistance of measurement result.**  **Proposal 27: Regarding UE reporting for network side models, RAN4 can wait for a more concrete proposal to be made in other WGs, e.g., RAN1 or RAN2.**  **Proposal 28: Regarding different types of performance monitoring for UE-sided model, RAN4 requirement may be provided as follows:**   * **For Type 1, Option 1 (NW-side performance monitoring): From RRM requirements perspective, it hasn’t fundamental difference from the legacy measurement scheme.** * **For Type 1, Option 2 (UE-assisted performance monitoring): New RRM requirements on the performance metric.** * **For Type 2 (UE-side performance monitoring): Wait for further clarification in RAN1.** |
| [**R4-2412250**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412250.zip) | vivo | **Observation 1: There is a non-negligible performance degradation on beam prediction accuracy when considering measurement error modelling in training/validation data.**  **Observation 2: 1dB margin could bring a noticeable gain on the prediction accuracy.**  **Observation 3: Training/validation dataset inconsistency due to the measurement error slightly impact the prediction accuracy.**  **Observation 4: There is no large performance gap on the RSRP accuracy between the following two definitions of RSRP difference**   * ***Alt1: RSRP difference only pertains to the case if AI/ML predicted beam ID is consistent with corresponding genie beam, or*** * ***Alt2: RSRP difference between predicted Top-K beam with corresponding Top-K genie beam***   **Observation 5: It can be observed that when beam prediction accuracy is high, the corresponding RSRP accuracy is also within reasonable range**  **Proposal 1: RAN4 to use beam prediction accuracy as the baseline of KPI and whether RSRP accuracy can be the KPI depends on if the RSRP is mandatory to be reported in the report content**  **Observation 6: Whether a high SNR is needed or not depends on the KPI selection and the requirement.**  **Proposal 2: Before determining the certain SNR level, the simulations of whether high SNR level is necessary and the impact due to Rx beamforming gain may be needed.**  **Observation 7: The multiple-AoA test setup depends on the test framework progress.**  **Observation 8: “Same” channel condition meaning is diverse, e.g., same channel coefficient, same channel large/small scale parameter, etc.**  **Proposal 3: Same channel conditions are needed when testing to ensure the RSRP values are unchanging and comparable.**  **Proposal 4: RAN4 to consider the maximum number of set B and set A Tx beams that test system should be able to emulate: (64,128).**  **Proposal 5: RAN4 to first study the test system framework and come back to discuss the number of AoAs after more progress are made.**  **Proposal 6: UE rotation is needed for time domain beam prediction. For spatial domain beam prediction case, there is no need to rotate the DUT in every single test. RAN4 to consider whether and how to rotate the UE for generalization tests.**  **Proposal 7: RAN4 to use CDL channel model as the baseline for the AI/ML based BM test.**  **Proposal 8: RAN4 to consider the framework of FR2 MIMO OTA test as the test framework for AI/ML based BM, by at least considering the following aspects:**   * **The number of DL Tx beams that need to be emulated** * **The number of OTA probes** * **The range of OTA probes deployment** * **The channel needs to be reconstructed** * **The rotation of DUT for emulating movements** * **The changing of the selected probe/probe power weight/coefficients running in the channel emulator**   **Proposal 9: UE rotation coordinates with the changing of the channel coefficient needs to be considered for BM case 2.**  **Proposal 10: For BM case 2, the probe power weight can be derived by considering:**   * **Different AOA/ZOA of clusters and different number of clusters** * **Different or same probe deployment (number and position of probes) for different DL Tx beams** * **Different probe power weight for different DL Tx beams**   **Observation 9: The number of probes can be greatly reduced by grouping different Tx beams together. Only one probe or one set of probes can emulate different Tx beams in one group.**  **Observation 10: Different test system with reduced cost and complexity can be introduced by using the Tx beam grouping method.**  **Proposal 11: RAN4 to consider methods to group Tx beams together to reduce the number of probes in the test system.** |
| [**R4-2412766**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412766.zip) | Huawei,HiSilicon | ***Proposal 1:*** Before investigating how to set up the test environment, the following questions have to be answered:   * How to ensure the consistency between model training and model testing   + Whether and how to define a common training/testing dataset.     - If define, how to capture the different UE implementations and UE behavior when constructing the dataset   + Whether UE is expected to collect training dataset and train the model before performing model performance test.     - If yes, how to resolve the test cost/time issue. * How to ensure that a UE can pass the test but perform poorly in the field, considering that some parameters used in the test set up which limit the model generalization may totally be different from that in real deployment.   ***Proposal 2:*** For test set up in AI-BM, taking the existing FR2 OTA test set up as baseline, any enhancements on top of which should be justified. |
| [**R4-2412994**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412994.zip) | Nokia | In this paper we share our views on potential RAN4 impacts from issues related to AI/ML enabled BM-Case1 and BM-Case2. Specifically, we cover following aspects:   * Ground truth reporting and measurement accuracy * Indicated/activated TCI states requirements * LCM related requirements * Generalization and test coverage related aspects * Testing setup   In the paper, the following Observations and Proposals were made:  **Observation 1:** Existing principles of testing UE-reported RSRP accuracy based on setup parameters and measurement uncertainties allow for very wide range of valid RSRP values. This makes testing for verification of AI/ML-enabled RSRP/best beam prediction a very futile exercise.  **Observation 2:** No significant improvement in L1-RSRP measurement accuracy is observed while changing SINR side condition from -3 to 0 dB.  **Observation 3:** If the requirement on predicted RSRP accuracy is based on the error-prone measured RSRP, the resulting absolute error of the prediction can be higher than in legacy requirements.  **Proposal 1:** **If reported RSRP measurements are used as an approximate of the ground truth, a tighter than legacy measurement accuracy should be considered for the definition of requirements for AI/ML BM.**  **Observation 4:** If both measurements and predictions are reported by the UE, there is no objective/external criteria to verify the independence of these reports.  **Proposal 2: The requirement based on RSRP measurements used as an approximate of the ground truth should not be the only ones specified for BM use cases.**  **Observation 5:** The behaviour of beam prediction functionalities can be tested without explicit definition of the ground truth by comparing the predictions to each other.  **Proposal 3: RAN4 to introduce additional requirements based on relative difference in between the predictions due to change in the conditions/inputs, i.e., without dependency on the measurement accuracy errors and uncertainties.**  **Observation 6:** When the TCI state is known, it is assumed that UE is aware of corresponding beam/RS, it has measured it recently, UE can switch to known TCI state faster than to an unknown TCI state.  **Proposal 4: RAN4 should consider relaxing known conditions for TCI states (Clause 8.10.2 of TS 38.133) for BM-Case1 and BM-Case2**  **Observation 7:** The delay considered for these operations necessary to activate the functionality depends on the use case (spatial beam prediction or temporal beam prediction) and should be carefully analysed.  **Observation 8:** If an LCM action is required and it is not taken in a timely manner, the performance for AI/ML enabled BM use case may be degraded to undesirable level.  **Proposal 5: Core requirements should be considered to limit latency of LCM actions typical for AI/ML enabled BM-Case1 and BM-Case2.**  **Proposal 6: Activation, Deactivation, Switching, Fallback to non-AI operation and Performance monitoring should be prioritized for defining RRM requirements for LCM AI/ML BM use case.**  **Observation 9:** A UE model/functionality, if tested in different generalization scenarios/configuration, would be considered not overfitted to a certain test environment.  **Proposal 7: Testing of generalization aspects should also be considered to avoid overfitting of UE models to the test environment.**  **Proposal 8: For the verification/testing of generalization related aspects in RAN4 for AI/ML BM, RAN4 should define different scenarios based on parameters listed in the Table below.**  Parameters for Generalization Scenarios   |  |  | | --- | --- | | **Parameters** | **Description** | | Propagation Model | AWGN/CDL/Uma/UMi | | SINR | Good / Bad Radio conditions | | UE Speed | Slow / Medium / Fast | | Channel propagation conditions | LOS/NLOS |   **Proposal 9: For the verification/testing of scalability related aspects in RAN4 for AI/ML BM, RAN4 should define different scenarios based on parameters listed in the Table below.**  Parameters for Scalability Scenarios   |  |  | | --- | --- | | **Parameters** | **Description** | | gNB antenna configurations | gNB antenna array 2x4/4x8/8x16 | | Variable number of Set B beams | Set B 16/32/64 beams | | UE Rx beams | UE Rx beams 4/8 beams per panel |   **Proposal 10: RAN4 needs to design a new metric, indicative of generalization capabilities of model/functionality for AI/ML BM, to verify the generalization performance of the model/functionality in different scenarios.**  **Observation 10:** With L1-RSRP being input to the AI/ML model/functionality for BM use case, it is redundant to test how the DUT measures the L1-RSRP during the AI/ML enabled BM conformance test and hence AWGN channel model may be sufficient for testing setup of this use case.  **Observation 11:** Use of complex testing setups for the conformance testing of AI/ML enabled BM use case will not add any significant value over the simpler testing setups but will increase the test complexity for sure.  **Proposal 11: As a starting point, RAN4 should focus on simpler test setups (e.g., AWGN based) for the conformance testing of AI/ML enabled beam management use case and not on the complex ones.**  **Observation 12:** A combination of AWGN based test system and UE rotation can be a potential candidate for the conformance testing of AI/ML based BM use case.  **Proposal 12: RAN4 to discuss and explore the feasibility of performing beam sweeping in multiple iterations during the conformance testing of AI/ML based BM use case.** |
| [**R4-2413038**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413038.zip) | ZTECorporation,Sanechips | **Observation 1: For RSRP accuracy prediction, the RSRP difference is mainly aim to the same beam between the predicted RSRP value and the measured or legacy RSRP values.**  **Observation 2: For legacy L1-RSRP accuracy test metric, ideal RSRP is a range not a single value which has the lower and upper bounds.**  **Observation 3: The relative accuracy is the measured L1-RSRP difference between two RSs.**  **Proposal 1: The predicted L1-RSRP difference is similar to the legacy, the absolute RSRP accuracy and the relative RSRP accuracy can be:**  ***Absolute RSRP accuracy= predicted L1-RSRP of beam index i – ideal measured L1-RSRP of beam index i***  ***Relative RSRP accuracy= predicted L1-RSRP of beam index i – predicted L1-RSRP of beam index n, where the beam index n owns the largest predicted value.***  **Observation 4: TE will verify whether the predicted strongest beam ID is the same as strongest measured beam or legacy beam ID. If the strongest beam ID is same as legacy, that means the AI/ML method is better. Otherwise, the test fails.**  **Proposal 2: Option 2 shall be considered as the performance metrics for beam prediction.**  **Observation 5: For option 3, there are two different understandings and the different understanding will cause the different outcomes.**  **Observation 6: Alt.1 does not emphasize whether the strongest RSRP of predicted beam ID is the same with the reference beam ID or not.**  **Observation 7: Alt.1 may have worse average RSRP prediction accuracy level even if the UE passes the test.**  **Proposal 3: The possible worse average RSRP prediction accuracy shall be considered if option 3 Alt.1 is chosen to be one of the performance metrics.**  **Proposal 4: RAN4 shall discuss which alternative shall be considered for option 3 and the understanding for option 3 shall be aligned.**  **Proposal 5: The description of option 3 shall be refined if Alt.2 is considered and it can be refined as:**  ***The reference RSRP value of the beam ID corresponding to the maximum predicted RSRP value shall larger than the strongest reference RSRP value minus x dB.***  **Observation 8: AWGN model does not have the spatial correlation.**  **Proposal 6: For AI beam prediction, AWGN model shall not be used and fading channel shall be considered.** |
| [**R4-2413168**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413168.zip) | Qualcomm Incorporated | **Observation 1: An effective test case for beam prediction use case BM-Case1 requires the following:**   * **Sufficient randomness and variation in time and spatial domain of L1-RSRP has to be emulated in the test** * **Support emulation of DL Tx beam sweeping with enough number of Tx beams in Set B and Set A**   **Observation 2: RAN4 L1-RSRP and other measurement test configurations support only:**   * **Deterministic and static power configuration on each AoA** * **Emulation of received signals from up to 2 AoAs**   **MIMO OTA testing environment defined in Rel-17 support only:**   * **Emulation of CDL channel with fixed DL beam to the channel** * **Emulation of received signals from up to 6 AoAs/probes, and the coverage (in terms of area on the sphere) by the 6 probes is very limited and sparse based on TR 38.827 Table 6.2.3-1, most of them are in Theta [0,30] and Phi [0,120], only one “mirror” point at negative Phi region.**   **Observation 3: The following conditions contribute to randomness and variation in time and spatial domain of L1-RSRP**   * **Propagation conditions as a function of (1) AoD of the Tx beam (2) AoA of the Rx beam (3) fading condition, e.g., a CDL channel** * **Tx beamforming gain on the AoDs in the propagation conditions** * **UE movement**   **Proposal 1: DL Tx beam sweeping for BM-Case 1 test can be emulated by the time-varying input power to the CDL channel model emulated on top of MIMO OTA test environment.**  **Proposal 2: Received power emulation can follow the formula below:**  **Where is Tx beamforming gain as a function of beam index and AOD, and channel gain (from CDL channel model) on each path, , is a function of AOD and AOA of the path.**  **Proposal 3: Tx beam sweep can be emulated based on the following formula by the probes on AoAk:**  **Where AoAk is the AoA of probe *k*, channel gain at time *t* denoted by sampled from CDL model, and Tx beam used in time *t* denoted by based on the beam sweeping RS transmission pattern. Note that the probe power configuration can be verified by comparing the probe power as a function of time and the distribution of .**  **Observation 4: UE received power when DL Tx beam is transmitting can be formulated as:**  **However, this quantity is unknown to TE since TE doesn’t have access to the .**  **Observation 5:** **One candidate resolution to ground truth availability is explained in the following. We can configure the test and TE channel emulator (fader) so that the channel from t0 to t1 is the same as channel from t1 to t2**  **During t0 to t1, only set B is transmitted; and in t = t1 to t2, set A is transmitted.**  **Proposal 5: The following issues should be considered when defining the beam prediction accuracy requirements**   * **Consistency between training and testing data (from the perspectives of beams in Set B and Set A, physical characteristics of gNB antenna etc.) should be guaranteed by signaling conveyed to UE.** * **The impact of size and composition of Set B and Set A on accuracy requirement.**   **Observation 6: Option 2 (beam prediction accuracy) does not distinguish between the following two scenarios**   * **Scenario A: UE’s top-K predicted beams are far from the top-1 strongest beam.** * **Scenario B: UE’s top-K predicted beams are adjacent to but don’t include the top-1 strongest beam**   **Proposal 6: RAN4 prioritizes option 3 for the performance evaluation of BM-Case 1 and BM-Case 2.**   * **Value of X is FFS.** * **Note: X can be defined based on RAN4 study.**   **Observation 7: Following steps can be considered to test option 3 for the performance evaluation of BM-Case 1 and BM-Case 2.**   * **Step 1:**    + **TE transmits set B beams.**   + **UE reports top K predicted beams of set A based on measurement of set B beams.** * **Step 2:**    + **TE randomizes set A beams and then transmits them (i.e., during time t1, UE does not transmit with beam shape A1 and UE transmits using the shape of a beam whose ID is randomly selected from set A).**   + **UE reports measured RSRP of all beams of set A.**   + **TE compares measured RSRP of strongest beam and maximum RSRP of top K predicted beam.** |

## Open issues summary

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

1. RSRP Prediction accuracy definition
2. Issues related to ground truth
3. Channel model in tests
4. Test setup needs/requirements
5. OTA test setup for beam management
6. Measurement error impact evaluation
7. UE reporting for network side models
8. Tx beam prediction and TCI state related issue

### Sub-topic 2-1

*RSRP Prediction accuracy*

The definition of the prediction accuracy was discussed in the previous meeting but there were no agreements yet.

**Issue 2-1: RSRP Prediction accuracy**

* Proposals
  + Option 1: The predicted L1-RSRP difference is similar to the legacy, the absolute RSRP accuracy and the relative RSRP accuracy can be:
    - Absolute RSRP accuracy= predicted L1-RSRP of beam index i – ideal measured L1-RSRP of beam index I (ground truth as agreed in RAN4#111)
    - Relative RSRP accuracy= predicted L1-RSRP of beam index i – predicted L1-RSRP of beam index n, where the beam index n owns the largest predicted value.
    - Accuracy requirements should apply to all predicted beams that satisfy the predefined side conditions, including SNR.
  + Option 2: others
* Recommended WF
  + Option 1

to be discussed whether any refinement is needed to the proposed Option 1

### Sub-topic 2-2

*Issues related to ground truth*

It was agreed in RAN4#111 that the ground truth for the predicted RSRP is as below.

* The ground truth for the predicted RSRP is the ideal measurement of RSRP on the predicted Tx beam
  + In RAN4, the ground truth is the approximate as the reported RSRP measurement under the certain SNR on the predicted Tx beam
    - FFS on SNR level to ensure that SNR is high enough for sufficient accuracy of reported RSRP
    - FFS on impact of multiple-AoA test setup
    - FFS on the channel condition
    - FFS on whether the ground truth will be changing or not
  + Other solutions are not precluded

The agreement contains many FFS and there are many practical issues related to tests that need further discussion/consideration.

Multiple companies brought up several issues that can be discussed separate, however, they are grouped under the same issue.

**Issue 2-2: Reported measurements and ground truth related issues**

* Issues raised:
  + Option 1: UE will report predicted RSRP/beam at T1 and measured best beam as ground truth at T2 to TE, it’s easy for UE to pass the test by cheating.Option 2: Ground truth cannot be determined
  + Option 2: Suggest to use RSRP delta CDF-based simulation to determine the appropriate SNR level for deriving ground truth.
  + Option 3: If reported RSRP measurements are used as an approximate of the ground truth, a tighter than legacy measurement accuracy should be considered for the definition of requirements for AI/ML BM.
    - Define tighter RSRP accuracy with a high SNR side condition?
  + Option 4: One candidate resolution to ground truth availability is explained in the following. We can configure the test and TE channel emulator (fader) so that the channel from t0 to t1 is the same as channel from t1 to t2
    - During t0 to t1, only set B is transmitted; and in t = t1 to t2, set A is transmitted.
  + Option 5: Other issues
* Recommended WF
  + To be discussed

### Sub-topic 2-3

*Channel models in tests*

The issue about the channel model to be used in prediction tests has been discussed in previous meeting without consensus. Multiple contributions discuss the channel model needs and issues in emulation. The channel model is an important component of the test setup.

**Issue 2-3: Channel models**

* Proposals
  + Option 1: CDL
  + Option 2: TDL
  + Option 3: AWGN
  + Uma/Umi
  + Option 4: simplified CDL
    - CDL channel model with a small number of clusters. CDL simplification needs to be further discussed
  + Option 5: others
* Recommended WF
  + Option 4

To be discussed whether more details about how to simplify CDL (or how to apply spatial filtering) can be agreed

### Sub-topic 2-4

*Test setup needs/requirements*

A list of test setup requirements was discussed in the previous meeting and some agreements were made as below:

maximum number of set B Tx beams that test system should be able to emulate: [8-16]

maximum number of set A Tx beams that test system should be able to emulate: [64-128]

FFS on AoAs

UE rotation during the test: FFS

UE rotation/repositioning between different tests: Yes

More discussion is needed on other parameters

**Issue 2-4: Test setup needs**

* Proposals
  + Option 1:
    - set B Tx beams: baseline 8; set A Tx beams: baseline 64
      * Does the TE need to broadcast all Tx beams in the same test?
    - AoA
    - AoD
    - UE rotation during the test
    - multiple TRP (or reflections from another direction)
  + Option 2:
    - other parameters
* Recommended WF
  + To be discussed

### Sub-topic 2-5

*OTA test setup for beam management*

Some companies provided analysis on the available test setups, pros/cons and capabilities. To progress the discussion, a more in depth analysis of the test setup capabilities is needed

**Issue 2-5: OTA Test setup for beam management**

* Proposals
  + Option 1: Take FR2 MIMO OTA test setup as baseline for further study analysis
  + Option 2: Multi-AoA IFF based system with a sparse probe layout as baseline
  + Option 3: existing IFF or enhance IFF test systems with single AoA for AI/ML BM test
  + Option 4: existing RRM test setup with 2 AoA
  + Option 5: continue discussion, discuss how to handle the test setup study
  + Option 6: others
* Recommended WF
  + To be discussed

It should be discussed how to continue the study on the test setup, whether a plenary recommendation can be made to continue the study under the FR2 OTA enhancements or in this WI or in some other SI

### Sub-topic 2-6

*Measurement error impact evaluation*

Multiple companies brought up the need to evaluate the impact that the UE measurement error and the error in the training data have on inference accuracy

**Issue 2-6: Measurement error impact**

* Proposals
  + Option 1: RAN4 should study the impact of measurement error, companies should bring proposals for the next meeting on how to proceed with such a study
  + Option 2: RAN4 should postpone the discussion on the impact of the measurement error for now
  + Option 3: No need to evaluate the measurement error impact
* Recommended WF
  + To be discussed

### Sub-topic 2-7

*UE reporting for network side models*

One company brought up a possible need for RAN4 to introduce different reporting schemes to help train the network side models.

**Issue 2-7: UE reporting for network side models**

* Proposals
  + Option 1: RAN4 shall introduce the necessary core requirement on supporting data collection for NW-side AI/ML model inference/training (for BM-Case1 & 2), by considering:
    - Potential enhancement on L1 measurement/reporting for inference: e.g., beam reporting for more than 4 beam in L1 signaling, and overhead reduction;
    - Potential enhancement on MDT-based measurement/reporting for training.
  + Option 2: RAN4 cannot introduce new reporting schemes, proposal should be made in another WG
  + Option 3: others
* Recommended WF
  + To be discussed

### Sub-topic 2-8

*Beam prediction and TCI state related issues*

Some companies brought up the issue that the UE will predict the gNB Tx beam but it might know the UE Rx beam. This could affect the known/unknown TCI state conditions

**Issue 2-8: Beam prediction and TCI state related issues**

* Proposals
  + Option 1: Tx beam prediction does not imply that UE knows the best UE Rx beam to receive the predicted beam
    - TCI state for the predicted beam is unknown unless it has been explicitly measured within 1280ms
  + Option 2: UE knows the Rx beam paired with the predicted Tx beam
    - FFS whether the TCI state associated with a predicted Tx beam is known or unknown
  + Option 3: UE knows the Rx beam paired with the predicted Tx beam
    - TCI state should be considered known , FFS whether side conditions for known TCI state should be modified (relaxed) for this case
  + Option 6: others
* Recommended WF
  + To be discussed

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2411293**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411293.zip) | OPPO | **Proposal 1: For case1, RAN4 will not define positioning accuracy requirements in R19 WI.**  **Proposal 2: For case1(AI/ML assisted positioning),**  **- feasibility of using these intermediate results(output of AI/ML models with non-linear processing) for case 1 tests is not clear.**  **- how to get the label data for intermediate results(e.g. RSTD, identification of LoS/NLoS and other metrics that agreed in RAN1) and test these intermediate results is not clear.**  **Proposal 3: For case2a, should be treated with 2nd priority in RAN4, the feasibility of using these intermediate results(output of AI/ML models with non-linear processing) for case 2a tests is not clear.**  **Proposal 4: For case2b/3a/3b (cases without UE-side model), not necessary to test the Positioning model /functionality outputs.** |
| [**R4-2411342**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411342.zip) | CATT | **Proposal 1: RAN4 stick to the agreements achieved in RAN4#110bis, i.e., RAN4 should postpone the discussion until other WG conclude on defining a reporting scheme or not.**  **Proposal 2: RAN4 to postpone the discussions on defining performance requirements for supported reported metrics since it is possible to reuse the existing performance requirements.**  **Proposal 3: RAN4 postpone the discussions on defining performance requirements for CIR until RAN1 agrees to support it.**  **Observation 1: LOS/NLOS indicator is not introduced due to AI/ML positioning use cases and no performance requirements are defined for it.**  Proposal 4: RAN4 not to define performance requirements for LOS/NLOS indicator report unless strong motivation is identified. |
| [**R4-2411410**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411410.zip) | Apple | **Proposal 1: RAN4 to further discuss the feasibility and how to define requirements for Positioning accuracy for case 1. Positioning test data sets could be one option for testing this KPI. Privacy from UE side should also be considered.**  **Proposal 2: For Assisted AIML Positioning, the KPIs test metric (e.g., LOS/NLOS) needs to be considered for validating the positioning accuracy**  **Proposal 3: In AI/ML-based positioning, it is essential to investigate the performance requirements for the input parameters of the positioning model/functionality (e.g., measurement accuracy of RSRP, ToA, RSRPP, RSTD)**  **Proposal 4: RAN4 to study defining performance accuracy requirements for any reported UE measurements defined by other groups.**  **Proposal 5: RAN4 to define performance accuracy requirements for use case 3a for the measurements reported by gNB. The specifics and scope of the core and performance requirements for use case 3a are contingent on the outcomes of discussions within RAN1 and RAN2**  **Proposal 6: RAN4 to define performance accuracy requirements for measurements performed at gNB for use case 3b. The specifics and scope of the core and performance requirements for use case 3a are contingent on the outcomes of discussions within RAN1 and RAN2** |
| [**R4-2411627**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411627.zip) | Xiaomi | **Observation 1: The legacy PRS measurement requirements up to Rel18[5] are based on the path based measurement.**  ***Proposal 1: If RAN1 using sample-based measurement as model input for AI Pos, the existing accuracy requirements (e.g. PRS RSRPP) shall be restudied at least.***  **Observation 2: For AI model performance monitoring, the proper threshold used to justify AI model validity can be studied by RAN4.**  **Observation 3: For Case 1 if AI model performance monitoring handled by UE, the requirements on model evaluation period can be defined in RAN4.**  ***Proposal 2: RAN4 shall study the following necessary requirements for AI pos model monitoring. The specification impacts on TS38.133 or other 3GPP spec can be FFS also.***   * ***Monitoring hypothesis and threshold*** * ***Performance evaluation period*** |
| [**R4-2411682**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411682.zip) | NTT DOCOMO, INC. | **Proposal 1: No additional requirements for positioning accuracy for case 1. It can be ensured if training data is accurate.** |
| [**R4-2411981**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411981.zip) | CMCC | ***Proposal 1: for case 1, whether to define requirements can be further discussed based on RAN1/2 conclusion.***  ***Proposal 2: for case 2a, it is proposed that LOS/NLOS indicator can be treated as intermediate performance metric, and to define requirements for LOS/NLOS indicator.***  ***Proposal 3: for case 2a, it is proposed that DL RSTD and UE Rx-Tx time difference can be treated as intermediate performance metric.***  ***Proposal 4: for case 2b, it is proposed to discuss whether legacy measurement accuracy requirements for DL RSTD, UE Rx-Tx and PRS-RSRP can be reused or better performance can be specified.***  ***Proposal 5: for case 3a/3b, it is proposed to define requirements for the measurement reported by gNB, and legacy requirements for positioning measurement performed by gNB can be conaisdered as baseline.***   * ***There are legacy requirements for positioning measurement performed by gNB, i.e. gNB Rx-Tx time difference absolute accuracy in 13.2.2.2 of TS 38.133, and gNB SRS-RSRP in 13.3.2.2 of TS 38.133.***   ***Proposal 6: when specify performance reqirements for AI/ML based positioning, it is proposed to discuss whether and how to consider the impact due to different assumption, e.g.model-input Size Reduction, non-ideal label(s), etc.*** |
| [**R4-2412251**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412251.zip) | vivo | ***Observation 1: Potential report schemes need to be introduced in performance monitoring procedures according to RAN1 agreements and RAN4 requirements for delay/accuracy may need to be considered in different options.***  ***Observation 2: Delay/accuracy requirement for performance monitoring procedure in case 1 may need to be defined and more progress from RAN1/2 are needed for detailed further discussing.***  ***Observation 3: More progress for the monitoring procedure from other WGs are needed to consider whether and how to define requirement/test for the potential metrics in case 1.***  ***Observation 4: Existing LPP procedures related to Location Information Transfer are used for case 1 and the UE position, which is the output for case 1, is supported for reporting.***  ***Observation 5: PRUs input and ground truth location could be used for testing the AI model for case 1.***  ***Proposal 1: RAN4 to consider whether and how to define accuracy requirement and test for UE position in case 1.*** |
| [**R4-2412695**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412695.zip) | Ericsson | ***Proposal 1: RAN4 to define core requirement for use case 2a where RSTD measurement is the AI/ML model output.***  ***RAN4 to define core requirement for use case 2a where UE Rx-Tx time difference measurement is the AI/ML model output.***  ***Observation 1: Training is critical for AI/ML model inference and retraining is prominent aspect relating to LCM of an AI/ML model, it is important that RAN4 takes these aspects into account and discuss delay requirement for AI/ML model training for use case 2a based on the agreement in other WG.***  ***Observation 2: The existing requirements for the RSTD and UE Rx-Tx measurements may not be relevant for the RSTD and UE Rx-Tx measurements reported by UE as AI/ML model inference.***  ***RAN4 to at least start discussion on simulation assumptions to evaluate achievable accuracy of AI/ML based RSTD and UE Rx-Tx measurements.***  ***Proposal 2: RAN4 to define measurement requirements for path based and sample-based reporting for use case 2b.*** |
| [**R4-2412767**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412767.zip) | Huawei,HiSilicon | ***Proposal 1:*** For UE-sided direct AI/ML positioning (Case 1), positioning accuracy is not testable.  ***Observation 1:*** For UE-assisted positioning with LMF-side positioning (Case 2b), the relationship between measurement accuracy and positioning accuracy is unavailable, which has an impact on the test requirement definition.  ***Proposal 2:*** If RAN4 studies how to test Case 2b in AIML for positioning, the relationship between measurement accuracy provided by UE and the eventual positioning accuracy at LMF needs to be investigated firstly. |
| [**R4-2413039**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413039.zip) | ZTECorporation,Sanechips | **Observation 1: There is no need for RAN4 to consider model inference for case 2b since the AI/ML model is deployed at LMF side which the output UE location on the LMF side depends on the network implementation.**  **Proposal 1: RAN4 shall not define the accuracy requirements for case 2b, it is up to network implementation.**  **Observation 2: In legacy, there are two kinds of location request: UE-triggered location request and NW-triggered location request. The latest agreement contains the implicit location request which is the NW triggers the location request, otherwise no reporting scheme is needed.**  **Proposal 2: RAN4 shall wait for RAN1 agreements on defining reporting scheme.**  **Observation 3: For AI/ML based positioning, the main difference compared to the legacy is that the AI model resides within the LMF. The accuracy of position prediction conducted by the LMF depends on the measurements reported by UE.**  **Proposal 3: The intermediate features such as RSTD, RSRP etc. or some enhancements shall be considered.**  **Observation 4: From request to reporting time, there are two periods of time which are PRS measurement and model inference. If there is no limitation on these two periods, LMF would take a long time to receive an inference outcome or the intermediate features. It has the impact on the performance accuracy since the accuracy of model output will also decrease over time.**  **Proposal 4: RAN4 shall define the delay requirements from LMF requesting to UE reporting.**  **Proposal 5: ToA shall be the intermediate feature when studying the accuracy requirements for case 2a.**  **Observation 5: RAN1 just agreed at least LOS/NLOS indicator is supported for reporting.**  **Proposal 6: LOS/NLOS could be the intermediate feature to be reported and RAN4 shall consider how to define the requirements for LOS/NLOS.** |
| [**R4-2413291**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413291.zip) | Nokia | **Proposal 1: For AIML positioning, RAN4 to define the requirements on the measurements supported in legacy positioning (e.g., RSTD, PRS-RSRP, and PRS-RSRPP) and define the requirements on the new measurements (e.g., CIR, DP, and PDP) based on RAN1 progress.**  **Proposal 2: RAN4 requirements (e.g., in Cases 1, 2a) should be firstly defined with respect to single PFL.**  **Proposal 3: The consistency between training and inference should be achieved by having a requirement on consistent configuration (e.g., same PFL) when UE-sided models are used.**  **Proposal 4: RAN4 to study the requirements on a quality indicator for the ground-truth accuracy on positioning label in data collection.**  **Proposal 5: Positioning accuracy should be considered as the performance metric in Case 1 (UE-based positioning with UE-side model, direct AI/ML positioning).**  **Observation 1: Even when ground truth, i.e., accurately known position of the UE, for Case 1 AI/ML-based positioning is not available in the testing setup for absolute positioning accuracy calculation, the new functionality should not be left without any requirements or tests.**  **Proposal 6: RAN4 to consider relative positioning accuracy metric for Case 1, i.e., requirement on the possible change in AI/ML-based position due to a change/state of the environment/configuration.**  **Observation 2: Positioning accuracy can be verified based on the ground truth which may consist of the location points with known positioning co-ordinates (e.g., PRU or GNSS based).**  **Proposal 7: RAN4 should specify (i) the test mechanisms for positioning accuracy and (ii) the metric for accuracy performance in Case 1.**  **Observation 3: If an LCM action is required and it is not taken in a timely manner, the performance degradation for AI/ML enabled Positioning use case may be degraded to undesirable level.**  **Proposal 8: RAN4 to define the time latency limit on UE’s LCM actions indicated by network in Cases 1 and 2a.**  **Proposal 9: Activation, Deactivation, Switching, Fallback to non-AI operation, and Performance monitoring should be prioritized for defining RRM requirements for LCM AI/ML positioning use case.**  **Proposal 10: LOS/ NLOS indicator should be considered as an intermediate performance metric for case 2a (UE-assisted/LMF-based positioning with UE-side model, AI/ML-assisted positioning).**  **Observation 4: LOS/ NLOS as an intermediate metric/ KPI should be considered for assisted AI/ML positioning. LOS/ NLOS metric/ KPI validation requires labelled data.**  **Proposal 11: LOS/ NLOS metric/ KPI labelled data can be generated based on the ground truth extraction using the legacy methods with the help of a well calibrated device capable of precisely indicating the ratio of received LOS/NLOS signals.** |

## Open issues summary

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

1. Requirements for case 1
2. Requirements for case 2a
3. Requirements for reported metrics for case 2a
4. Requirements for case 2b
5. Requirements for reported metrics for case 2b
6. Requirements for reported metrics for case 3a/3b

### Sub-topic 3-1

*Requirements for case 1*

In the previous meeting it was discussed whether to introduce requirements for case 1 or not but no agreement was reached.

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

* Proposals
  + Option 1: RAN4 should not define requirements for case 1 because positioning accuracy is not testable
  + Option 2: RAN4 to further study defining requirements for case 1, companies are invited to bring further analysis:
    - how to establish the ground truth ( what should the UE report)
    - how to define accuracy requirements
    - what are the inputs for the AI/ML model
    - what side conditions to consider
  + Option 3: RAN4 should study how to define relative requirements (e.g. how the reported position changes based on changing inputs)
  + Option 4: Others
* Recommended WF

To be discussed

If Option 2 is agreed, other aspects to be studied can be discussed.

### Sub-topic 3-2

*Requirements for case 2a*

Some companies are proposing to define requirements for case 2a

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

* Proposals
  + Option 1: RAN4 should not define any positioning accuracy requirements because positioning is LMF based
  + Option 2: RAN4 should continue to study how to define requirements on the positioning accuracy of the LMF
  + Option 3: others
* Recommended WF
  + Option 1

### Sub-topic 3-3

*Requirements for reported metrics for case 2a*

Even if requirements for the positioning accuracy computed by the LMF are not defined for case 2a, some companies are proposing to introduce requirements on the metrics reported by the UE as inputs to the LMF positioning model.

**Issue 3-3: Reported metrics for case 2a**

* Proposals
  + Option 1: RAN4 to define requirements for LOS/NLOS reporting (reported metric agreed by RAN1)
    - companies to bring analysis on how such requirements can be defined and tested
  + Option 2: RAN4 to continue studying how to define requirements for other reported metrics (RSTD, UE Rx-Tx, etc)
    - Companies to bring analysis on how to define requirements and tests
  + Option 3: Defining requirements for reported metrics is not feasible and should not be further discussed
  + Option 4: Wait for other groups to agree on the reporting framework before continuing this discussion.
* Recommended WF
  + To be discussed, options are not exclusive

### Sub-topic 3-4

*Requirements for case 2b*

**Issue 3-4: Requirements for case 2b**

* Proposals
  + Option 1: RAN4 should not define any positioning accuracy requirements because positioning is LMF based
  + Option 2: RAN4 to further discuss how to define requirements for the LMF based positioning accuracy (LMF based positioning model)
  + Option 3: others
* Recommended WF
  + Option 1

### Sub-topic 3-5

*Requirements for reported metrics for case 2b*

Case 2b is based on UE reports of different metrics which are not computed based on AI/ML models. Existing reported metrics (e.g. RSTD) can also be reused

**Issue 3-5: Reported metrics for case 2b**

* Proposals
  + Option 1: RAN4 to discuss requirements for case 2b only if new reported metrics are introduced by other groups
    - existing requirements for legacy reported metrics will not be revisited
  + Option 2: RAN4 should revisit the existing requirements for legacy reported metrics
    - Companies to bring analysis on what can be improved(what has changed) for existing requirements
  + Option 3: Others
* Recommended WF
  + Option 1

### Sub-topic 3-6

*Requirements for reported metrics for case 3a/3b*

It was agreed not to define positioning accuracy requirements for case 3a/3b, however, some companies are proposing to define requirements for the metrics reported by gNBs which are inputs to the positioning models

**Issue 3-6: Reported metrics for case 3a/3b**

* Proposals
  + Option 1: RAN4 to further study how to define requirements for reported metrics for case 3a/3b
    - companies should bring concrete analysis on how the requirements could be defined and tested
  + Option 2: RAN4 should not define any requirements for gNB reported metrics
  + Option 3: Others
* Recommended WF
  + To be discussed

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2411178**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411178.zip) | Korea Testing Laboratory | **Observation 1: The interpretability, explainability, documentation, and traceability provided by legacy test cases should also be applied to AI/ML use cases, in accordance with the proficiency testing and interlaboratory comparison requirements outlined in ISO/IEC 17025.**  **Observation 2: A dataset needs to consider the various kinds of environmental scenarios since CSI is highly dependent on the propagation environment.**  ***Proposal 1: The detail parameters for Reference encoder and Reference decoder shall be specified (in TR, WF, etc) or captured in the specifications. (Fully standardized reference model (structure + parameters))***  ***Proposal 2: For the conformance testing, a dataset shall be specified (in TR, WF, etc) or captured in the specifications.*** |
| [**R4-2411294**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411294.zip) | OPPO | **Proposal 1: Principles to define test decoder(s)**   * **to meet the minimum performance requirement in RAN4 tests** * **to be a simple design**   **Proposal 2: Steps to determine test model(s):**   * **Setp1: Determine the test condition(high priority)** * **Step2: Determine the test data**   **Step 2-1 (high priority): consider how to set up the test data, how to determine the test date generation method**  **Step 2-2: consider how to collect datasets from different companies**  **Step 2-3: consider whether/how to merge datasets from different companies to form a test model data set**   * **Step3: Determine the test model structure**   **Step 3-1(high priority): consider how to set up the model structure**  **Step 3-3: consider how to merge and compromise different inputs on the proposed test model structures from different companies, to form a test model structure**   * **Step4: Determine the test model parameter**   **Step 4-1 (high priority): determine the test model parameters based on the test dataset and test model structure**  **Step 4-2: consider how to merge different sets of test model parameters to obtain a set of specified test model parameters**  **Proposal 3: Not necessary to spend a lot of time on discussing and aligning model hyperparameters and training methods.**  **Proposal 4: Regarding the test condition, e.g. test configurations or baseline scenarios, following parameters and candidate values should be considered and determined:**   |  |  |  | | --- | --- | --- | | Parameter | | Value | | Duplex, Waveform | | FDD, OFDM | | Multiple access | | OFDMA | | Scenario | | Dense Urban (Macro only) | | Frequency Range | | FR1 only, [2GHz, 4GHz] | | Inter-BS distance | | 200m | | Channel model | | According to TR 38.901 | | Antenna setup and port layouts at gNB | | Companies need to report which option(s) are used between  - 32 ports: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ | | Antenna setup and port layouts at UE | | 4RX: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ for (rank 1-4) | | BS Tx power | | 44dBm for 20MHz | | BS antenna height | | 25m | | UE antenna height & gain | | Follow TR36.873 | | UE receiver noise figure | | 9dB | | Numerology | Slot/non-slot | 14 OFDM symbol slot | | SCS | Baseline: 15kHz for 2GHz;  Optional: 30kHz for 4GHz | | Simulation bandwidth | | Baseline: 10 MHz for 15kHz  Optional: 20 MHz for 30kHz | | Frame structure | | Slot Format 0 (all downlink) for all slots | | MIMO scheme | | SU-MIMO | | MIMO layers | | Baseline: 1  Optional: 2 | | CSI feedback | | Feedback assumption at least for baseline scheme  -CSI feedback periodicity (full CSI feedback): 5 ms (baseline)  -Scheduling delay (from CSI feedback to time to apply in scheduling): 4 ms | | Traffic load (Resource utilization) | | Baseline: 50%  Optional: 20/70% | | UE distribution | | CSI compression: 80% indoor (3 km/h), 20% outdoor (30 km/h) | | UE receiver | | MMSE-IRC as the baseline receiver | | Feedback assumption | | Realistic | | Channel estimation | | Realistic or ideal channel estimation |   **Proposal 5: To determine a test data, link level test data should be utilized. Following parameters and candidate values should be aligned:**   |  |  | | --- | --- | | **Parameter** | Value | | Channel model | CDL-C as baseline | | Delay spread | 30ns | | Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline. |   Note: test data should be obtained from link level channel model, training data could be obtained from system level channel model, e.g. UMa channel  **Proposal 6: To determine a test model structure, following parameters and candidate values should be considered:**   |  |  |  | | --- | --- | --- | |  |  |  | | Model architecture parameters | Model type | CNN as baseline, e.g., Resnet based structure | | Model architecture | [TBD],  The agreed reference model structure (CNN) can be considered as starting point | | Quantization method for the encoder output | Scalar quantization as baseline, e.g. 2bit | | Encoder-decoder interface | - CSI feedback overhead A: 64 bits (small payload)  - CSI feedback overhead B: 112 bits (medium payload) | | Fixed point representation | floating point | | Format of input to encoder/output of decoder | Raw channel or CSI eigenvector |   **Proposal 7: Update the step 1-4 in the flowchart that discussed in RAN4#110bis meeting as below:**    **Proposal 8: For both CSI compression and CSI prediction, model/functionality input (CSI-RS measurement accuracy) and output (associated throughput) related tests should be supported.**  **Proposal 9: 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 10: Reuse the legacy PMI requirement (compare to random precoding) as a baseline test, and let other options/proposals with higher performance requirements be further studied.**  **Proposal 11: In R19, static test scenarios and configurations should be considered first. After having feasible testing cases for static configurations, then further consider whether to introduce non-static testing scenarios and configurations.**  **Proposal 12: Besides the TDL channel based tests, CSI-related tests can be conducted under CDL channel [or other more practical channel conditions] to check a relatively generalized performance.**  **Observation 1: FLOPs, Number of parameters, model size for CNN based CSI encoder and CNN based CSI decoder are evaluated as below:**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | CSI feedback bits | FLOPs | | Num. of Trainable Parameters | | Model Size | | | Encoder | Decoder | Encoder | Decoder | Encoder | Decoder | | 64 bits | 94.1M | 1481.6M | 0.13M | 1.80M | 638KB | 7.03MB | | 112 bits | 94.2M | 1481.7M | 0.14M | 1.82M | 677KB | 7.11MB |   **Observation 2: Based on CNN based CSI encoder and CNN based CSI decoder, performance evaluation results are shown as below(train by UMa channel data as agreed in the last meeting, test by different channel model)**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | CSI feedback bits | Training dataset | Test dataset (SGCS) | | | | | TDL-A30 | **CDL-A30** | CDL-C30 | **UMa** | | 64 bits | UMa | 0.759 | **0.705** | 0.603 | **0.725** | | 112 bits | UMa | 0.784 | **0.768** | 0.645 | **0.782** |   **Observation 3: In comparison to the CNN-based CSI encoder, there's a substantial decrease in FLOPs in the MLP-based CSI encoder, dropping from 94M to a mere 0.22M.**  **Proposal 13: For the reference CSI encoder, a simplified MLP-based CSI encoder could lower the challenges of aligning and implementing standardized CSI encoders across different companies, without compromising the CSI compression feedback performance.** |
| [**R4-2411343**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411343.zip) | CATT | **Observation 1: The SGCS of CNN based CSI compression is 75% with 64-bit feedback and 128 kernels in the decoder.**    **Figure1. Performance of CNN based CSI compression**  **Observation 2: 4%~6% SGCS degradation is observed when the number of kernels in the decoder decreases from 128 to 64.**    **Figure2. (a) Complexity of the decoder (b) Size of the decoder**  **Observation 3: Significant decreases are observed in complexity and number of parameters when the number of kernels in the decoder decreases from 128 to 64.**  **Proposal 1: Fewer kernels can be considered if RAN4 agree to define test decoders.**  **Proposal 2: It is not the first priority to define reference encoders, including backbone and structure.**  **Observation 4: The length of latent message is decided by encoder output which affects the implementation of decoders.**  **Proposal 3: RAN4 specify the length of latent messages instead of bit stream.**  **Proposal 4: Quantization scheme adopted by UE needs to be shared to TE. Two solutions can be considered:**  **(1) Specify a quantization scheme in test cases, or**  **(2) Follow RAN1 design** **on quantization scheme matching, if defined.**  **Proposal 5: UE under tests needs to report the input data of encoder to TE for performance evaluation.**    **Figure4. A high-level procedure of the feasibility study on Option 3 and 4**  **Proposal 6: Companies to run simulations with their own trained encoder and the chosen decoder and report the results for alignment. If results are not aligned, RAN4 to find a solution to specify channel characteristics, the following methods can be considered:**   1. **Capture the channel generation method into spec, and/or** 2. **Define datasets based on test scenarios and configurations, and/or** 3. **Other solutions, e.g., reference encoder/decoder.**   **Proposal 7: RAN4 not to discuss the TE verification/validation.** |
| [**R4-2411411**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411411.zip) | Apple | **Observation 1: Simulation results comparing SGCS between AI/ML and e-typeII are captured below:**   |  |  | | --- | --- | | **SGCS** | **Encoder/Decoder** | | **0.78** | AI/ML (64 bits) | | **0.747** | e-type II (configuration 1, 62 bits) |   **Table 3. Simulation results for AI/ML and e-typeII codebook**  **Proposal 1: For aiding the performance alignment of the simulation results reported by each company, each company should list the training parameters used for the encoder-decoder pair. An example of training parameters is shown below:**   * **Optimizer** * **Batch size** * **Loss function** * **Stopping Criterion** * **Max Epochs** * **Learning rate** * **Dataset size** * **Quantization of training data** * **Channel type** * **Weight initialization**   **Additionally, data sharing (Step 8) could be employed to aid misalignment issues:**   * **Company A trains its encoder/decoder with company’s B dataset.** * **Company B tests the encoder decoder pair with company’s A testing data**   **Proposal 2: RAN4 to employ the procedures shown in the flowchart for studying the feasibility analysis of test decoder Option 3.**  **Proposal 3: The following criteria should establish the feasibility of the test decoder:**   1. **The test decoder (along with the encoder) should deliver reasonable throughput performance gain compared to e-type II in ideal conditions. RAN4 to extend the feasibility analysis to higher ranks and different payload sizes to determine consistency of performance improvement. FFS on how to quantify the acceptable performance gain. (performance-complexity tradeoff)** 2. **All UE preferred encoders should be aligned (work well) with the chosen test decoder**   **Proposal 4: To determine if the performance aspect of the feasibility analysis in throughput is met, RAN4 should investigate an empirical model that maps SGCS to a throughput metric, such as spectral efficiency.**  **Proposal 5: The following principles should guide the establishment of the alignment aspect of the feasibility analysis (steps 6-7):**   * **A well-trained UE encoder with the same or higher complexity compared to the Ref encoder that the test decoder was trained with, should have at least similar performance with the reference performance. (in throughput or SGCS)** * **FFS on how to define a criterion for the similarity/dissimilarity of all companies' performance results (with UE preferred encoders) with respect to the reference performance, considering UE encoder complexity as well.**   **Proposal 6: RAN4 to define a criterion to select a test decoder among all the test decoders reported after performance alignment is achieved (step 5). The criterion could be based upon:**   * **Test decoder trained with the highest number of samples** * **Each company tests its pair of {Encoder, test decoder) with other companies testing data set. Chose the pair with best performance** * **Others**   **Proposal 7: According to the flowchart, RAN4 should first evaluate feasibility analysis with Option 1 and proceed to Option 2 if Option 1 does not produce a feasible test decoder**    **Figure 3. Flowchart of procedures for feasibility analysis of Option 3 Test decoder**  **Proposal 8: Consider the following aspects regarding the different conditions for testing generalization for CSI AI/ML use:**   * **Various antenna port layouts, e.g., (N1/N2/P) and/or antenna port numbers (e.g., 32 ports, 16 ports)** * **Various antenna spacings (e.g., 0.5 lambda, 0.8 lambda, etc)** * **Various antenna virtualization (TxRU mapping)** * **Various carrier frequencies and bands (e.g., 2GHz, 4.0GHz)** * **Various outdoor/indoor UE distributions for UMa/Umi** * **Various UE speeds.**   **Consider the following aspects regarding the scalability aspect for generalization testing for CSI AI/ML use:**     * **Various bandwidths (e.g., 20MHz, 50MHz) and/or frequency granularities, (e.g., size of subband), different layers** * **Various sizes of CSI feedback payloads** |
| [**R4-2411534**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411534.zip) | Rohde & Schwarz | **Observation 1: The computational complexity of the CNN-based (Resnet) test decoder model selected in the e-mail discussion following the RAN4#111 meeting is higher than originally reported (1,475 MFLOPs instead of 740 MFLOPs).**  **Observation 2: For the given scenario and latent space size, the computational complexity of the CNN-based (Resnet) test decoder can be significantly decreased by reducing the number of convolutional filters and/or Resnet blocks while suffering only moderate performance degradation.**  **Observation 3: For the given scenario and latent space size, both the CNN-based (Resnet) and MLP-based autoencoder models yield similar performance. The computational complexity of the MLP-based test decoder model is significantly lower than the CNN-based (Resnet) model proposed in the recent e-mail discussion.**  **Proposal 1: The computational complexity of the test decoder should be considered as an important KPI during the feasibility study and a potential future standardization.**  **Proposal 2: The potential for test decoder model complexity reductions should be considered during the feasibility study and a potential future standardization.**   |  |  | | --- | --- | | **Optimizer** | **Adam** | | Learning Rate | CNN-based model: 1e-4  MLP-based model: 1e-3 | | Learning Rate Adaptation | Reduce on plateau by factor 0.5 with patience of 50 epochs | | Loss Function | SGCS | | Batch Size | 512 | | No. Epochs | 200 |   Table 2 Training Parameters   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **No. of conv filters in decoder** | **No. of Resnet blocks in decoder** | **SGCS** | **MFLOPs** | **Remarks** | | 128 | 6 | 0.7650 | 1,475.4 | Original model proposed in e-mail discussion | | 4 | 0.7643 | 984.3 |  | | 2 | 0.7543 | 493.3 |  | | 1 | 0.7391 | 247.8 |  | | 64 | 6 | 0.7649 | 369.7 |  | | 1 | 0.7381 | 62.6 |  | | 32 | 6 | 0.7617 | 92.8 |  | | 1 | 0.7353 | 16.0 |  | | 16 | 6 | 0.7570 | 23.4 |  | | 1 | 0.7389 | 4.2 |  | | 8 | 6 | 0.7550 | 6.0 |  | | 1 | 0.7267 | 1.2 |  | | 4 | 6 | 0.7154 | 1.6 |  | | 1 | 0.7165 | 0.4 |  |   Table 3 Simulation results for CNN-based (Resnet) model |
| [**R4-2411628**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411628.zip) | Xiaomi | **Observation 1: The main difference between options is that option 2 provide latent information while option 1 provide fully decoder information.**  **Observation 2: Option 2 provide more limitation for performance generalization compared with option 1.**  **Proposal 1: For step 2, Adopt SLS based simulation assumption to generate training data set.**  **Proposal 2: For step 2, RAN4 needs to discuss whether each company train reference model by data set generated individually or mixed data set from multiple companies.**  **Proposal 3: For step 2, RAN4 needs to align the input training data format for data mixing or performance cross-checking.**    Fig.2 different encoder-decoder pair  **Proposal 4: For step 5, the candidate fully specified decoder should satisfy the condition that performance degradation due to different encoder-decoder pair will smaller than a threshold.** |
| [**R4-2411636**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411636.zip) | Samsung | *AI-CSI prediction:*  **Proposal 1:** For CSI prediction accuracy metric for inference, relative throughput (i.e., throughput by following predicted PMI over the baseline throughput) can be adopted, but   * FFS the definition of baseline throughput (as denominator):   + Option 1-1: Throughput achieved by following random PMI (with the same codebook used for the reported predicted PMI)   + Option 1-2: Throughput achieved by following UE’s last reported PMI given by UE measurement   **Proposal 2:** For relative throughput used as CSI prediction accuracy metric for inference, Option 1-2 (Throughput achieved by following UE’s last reported PMI given by UE measurement) is adopted for baseline throughput.  **Proposal 3:** In the study phase, no RAN4 discussion is needed on the testability issues for performance monitoring (including test method and test metrics to be used).  *AI-CSI compression:*  **Observation 1:** For the on-going evaluation (which is based on selected CNN-structure and trained over the company SLS-generated dataset under the specified scenario), it should be considered as a feasibility study on standardization procedure, i.e., for Step-6 and 7 in Samsung proposed flow-chart [R4-2405653]:    **Proposal 4:** For the on-going evaluation (which is based on selected CNN-structure and trained over the company SLS-generated dataset under the specified scenario), at least the following issues shall be studied:   * Issue-1: Whether or not the performance alignment can be achieved, with the selected CNN-structure and the company SLS-generated dataset under the specified scenario? * Issue-2: Performance degradation is acceptable if applying the trained model by company-A to the test dataset generated by company-B? * Issue-3: If the performance degradation in Issue-2 is not acceptable, could aggregating datasets from multiple companies solve the problem?   **Proposal 5:** The metric of FLOPs shall be adopted used to evaluate neural network complexity, and original CNN complexity values (47M and 740M for encoder/decoder respectively) are actually for MACs, which shall be corrected by FLOPs (9.39\*107 and 1.48\*109 for encoder/decoder respectively).  Table 4. Model training parameters   |  |  |  | | --- | --- | --- | |  | **Parameters** | **Values** | | **Model training** | Loss function | SGCS | | Training/validation datasets | Channel model for training:  UMa Channel based on TR 38.901 (details are given in the agreed WF [8])  Total dataset: 25K samples | | Hyperparameters | Learning rate = 0.001,  batch size = 200,  optimization algorithm = Adam | | **Others** | Cross-validation details | Dataset randomly split for training/validation |   Table 5. SGCS Performance for CNN-based Model Trained over Samsung-generated dataset   |  |  |  | | --- | --- | --- | | Encoder | Decoder | SGCS | | Back-bone: CNN  Number of model parameter: 125K  FLOPS: 9.39\*107  Depth: 6 Res blocks  Feature maps: 32 | Back-bone: CNN  Number of model parameter: 1.8M  FLOPS: 1.48\*109  Depth: 6 Res blocks  Feature maps: 128 | 0.76 |   **Proposal 6:** RAN4 shall use the SGCS performance reported from companies to confirm or disprove the performance alignment firstly, with SLS details to generate dataset to be further compared.  **Observation 2:** The neural network model trained in one dataset could suffer the substantial degradation when the model is tested over other datasets from other sources.  **Observation 3:** Possible reason for the above Observation 2 is (a) limited model generalization; (b) the training dataset size is not large enough; (c) the distribution for one dataset could be not representative enough.  **Proposal 7:** For CSI compression scenario, RAN4 can’t directly confirm the generalization performance for the company-wise model trained by using the dataset generated by SLS individually, even based on the detailed simulation parameters. |
| [**R4-2411711**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411711.zip) | MediaTek inc. | ***Table 2. Hyperparameters used for ResCNN model***   |  |  | | --- | --- | | ***Hyperparameters*** |  | | ***Number of samples in training dataset*** | ***700k*** | | ***Number of samples in testing dataset*** | ***160k*** | | ***Optimizer*** | ***Adam*** | | ***Learning rate*** | ***0.001*** | | ***Batch size*** | ***128*** | | ***Number of epochs*** | ***300*** | | ***Loss function*** | ***SCGS*** | | ***Drop rate*** | ***0.05*** |   ***Observation 1***: For the case of 2GHz and 1 layer, the SCGS of ResCNN model is 0.722 with 64 bits feedback. Also, the SCGS of eTypeII PC1 with 62 bits feedback and ResCNN model with 64 bits feedback is very close.  ***Observation 2***: FLOPs of ResCNN encoder is 93M and FLOPs of ResCNN decoder is 1480M.  ***Observation 3***: Parameters of ResCNN encoder is 0.127M and parameters of ResCNN decoder is 1.808M.  Table 3. SCGS for eTypeII and ResCNN   |  |  |  | | --- | --- | --- | |  | Overhead (bits) | SCGS | | eTypeII PC1 | 62 bits | 0.719 | | ResCNN model | 64 bits | 0.722 |   Table 4. Complexity of ResCNN AI/ML model   |  |  |  | | --- | --- | --- | |  | Parameters [M] | FLOPs [M] | | Encoder | 0.127 | 93 | | Decoder | 1.808 | 1480 | |
| [**R4-2411786**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411786.zip) | Moderator (Qualcomm Incorporated) | Paper containing the summary of the post meeting email discussion on the simulation assumptions for studying the feasibility of CSI testing Option 3 |
| [**R4-2411979**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2411979.zip) | CMCC | ***Proposal 1: for the feasibility study of option 3, based on the agreed system-level simulation parameters and proposed CNN model parameters in post-email email discussion, the evaluation results are shown as following:***  Table SGCS for CSI compression based on CNN model   |  |  | | --- | --- | |  | SGCS | | CNN | 0.792 |   ***Proposal 2: for CSI Prediction Accuracy metrics for inference, it is propose to use relative throughput, which is the throughput gain achieved with predicted PMI compared to random PMI.***  ***Proposal 3: for CSI compression and CSI prediction, it is proposed to use intermediate KPI, e.g. SGCS, as requirements/tests metrics for LCM.*** |
| [**R4-2412022**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412022.zip) | Nokia | **On test decoder derivation study/experiment:**   1. *tanh* activation function before the quantization layer was used for normalization and the following optimised hyperparameters were used during the training process as shown in Table 1 below.   Table 1:Optimized hyperparameters used during training.   |  |  | | --- | --- | | **Hyperparameter** | **Value** | | Optimizer | RMSProp (Root Mean Square Propagation) | | No of Epochs | 200 | | Batch Size | 1024 | | Learning rate | 10-4 | | Dropout rate | 0.4 | | L2 regularization coefficient (λ) | 10-2 | | Validation patience | 10 | | Loss function | SGCS |  1. RAN4 to consider the obtained SGCS metrics for the trained AI/ML encoder-decoder pair and eTypeII CSI feedback as follows:    1. Rel16 eTypeII baseline SGCS of Rank 1 eigenvectors at 2 GHz:   Table 2: CGSC performance of eTypeII CSI feedback   |  |  |  | | --- | --- | --- | | **Parameter Combination** | **Overhead Bits** | **Mean SGCS** | | 1 | 59 | 0.691 | | 2 | 88 | 0.732 | | Interpolation | 64 | 0.698 |  * 1. Trained AI/ML encoder-decoder pair SGCS performance:   Table 3: Trained AI/ML encoder-decoder pair SGCS performance   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Model** | **Overhead bits** | **Training SGCS** | **Testing SGCS** | **Parameters** | | CNN (NOKIA) | 64 | 0.644 | 0.643 | 1.93M |   **On derivation of test decoder and reference encoder for Option 4:**  RAN4 have not agreed on concrete criteria for the selection of the test decoder out of the whole set shared by the companies. Since the decoder + encoder architecture was fixed during the email discussion, complexity cannot be used as a selection criterion.  If Option 1 for test decoder derivation procedure is followed, additional criteria for the selection of a single decoder out of contributed are needed.   1. If fully specified test decoder (Option 3) needs to be selected from the models contributed by the companies (Option 1), then RAN4 to consider not only the best performance (in terms of SGCS) but also mean performance across all the companies and the performance of corresponding eType II CSI feedback as criteria.   Aggregated dataset {*V*} of channel eigenvectors (or channel realizations) can be used for more objective comparison/validation of encoders/decoders contributed by the companies.   1. RAN4 to use an aggregated dataset of channel realizations/eigenvectors {*V*} for comparison of the performance of encoders/decoders provided by the companies if it is necessary to select a single test decoder (can be considered as Option 1a). 2. To proceed with Option 2 an aggregated dataset of channel eigenvectors/realizations *V* and encodings *c*, i.e., {*V, c*}, should be created and then used for training, performance comparison, and selection of a single test decoder. 3. Selection of a single test decoder trained on aggregated dataset (Option 2) can be based on performance criteria, e.g., based on the best performance or closest to mean across companies.   Joint training of reference encoder with test decoder could provide too optimistic performance evaluation. It is not clear how reference encoder can be used for the derivation of the test decoder.   1. If a single reference decoder needs to be derived, a corresponding procedure needs to be agreed like for the test decoder, e.g.,    1. Option 1: RAN4 chooses one of the encoders trained by the companies based on the selected test encoder.    2. Option 2: A single reference decoder is trained based on the aggregated dataset.   **On partially specified decoder with Option 4:**  A reference encoder/decoder pair can be used as the source of the dataset {*V, c*} that specifies the feedback interface and ensures test decoder interoperability for Option 4.  Test repeatability in Option 4 can be addressed by developing a test decoder validation procedure including decoder performance specifications using the specified dataset.   1. Select a reference encoder/decoder pair to create a dataset for training Option 4 test decoders.   **On test decoder verification:**   1. For test decoder’s verification fully specified reference encoder should be used. 2. For test decoder’s verification, reference encoder may be implemented by UE vendors, TE vendors or a third-party implementation of UE sided operations including reference encoder. 3. For test decoder’s verification, the minimum performance target should be agreed and specified considering various results from different stakeholders.   **On metric for AI/ML-enabled CSI feedback requirements:**  The existing PMI/CSI reporting requirement is based on the relative throughput γ as the test metric. Type I single-panel codebook is easier to randomize to establish the reference PDSCH throughput (in the denominator of relative throughput γ) in comparison with Type II codebook or AI/ML-based compressed feedback.   1. Random precoding matrices from Type I single-panel codebook (mode 1 with appropriate number of layers) should be used a reference in the requirement on the relative throughput γ. 2. For AI/ML-based compressed CSI feedback, decoded eigenvectors of the channel matrix reported by the UE can be used for PDSCH precoding at SS.   We are expecting that in CSI prediction use-case legacy CSI reporting format will be used.   1. For AI/ML-based CSI prediction, CSI feedback predicted and reported by the UE in legacy format (e.g., eType II) can be used by SS for PDSCH precoding, i.e., like in legacy requirements.   The use of the same/comparable metric (relative throughput γ) for AI/ML based and existing legacy requirements with the same parameters/tests gives the way to compare the minimal level of performance.   1. RAN4 to ensure that relative throughput based on AI/ML CSI feedback with the same overhead is at least not worse than in legacy tests with the same parameters.   **On non-static scenarios:**  Non-static scenarios in RAN4 are needed to verify that there is no performance degradation while UE is autonomously switching in between different AI/ML models due to the change of radio conditions.   1. Non-static scenarios/conditions can include the change of radio conditions during the test, such as channel model (e.g., from indoor to outdoor) and/or its’s parameters (e.g., Doppler, delay spread). 2. RAN4 should define performance requirements for non-static scenarios/conditions provided that the functionality is not changing/switching during the test and internal model’s transitions are up to the UE.   **On generalization and test coverage:**  Generalization parameters like the overall scenario (LOS, NLOS, indoor, outdoor, etc.), SINR, UE speed, etc. are generally not known at the UE nor the gNB. And this must be configured at the TE.  Table 4: Parameters for Generalization Scenarios   |  |  | | --- | --- | | **Parameters** | **Description** | | UE Speed | Slow / Medium / Fast | | SINR | Good / Bad Radio conditions | | Outdoor / Indoor | Position of the UE | | LOS/NLOS |  | | Propagation Model | UMa / Umi |   Scalability parameters are generally known at the UE and the gNB and, typically, do not change during the active time of a UE in a certain cell.  Table 5: Parameters for Scalability Scenarios   |  |  | | --- | --- | | **Parameters** | **Description** | | Number of Antenna ports | (N1/N2/P) and/or antenna port numbers (e.g., 32 ports, 16 ports) | | Carrier Frequency | FDD, TDD at sub-band level | | Bandwidth | E.g., 10MHz, 20MHz |  1. For the verification/testing of generalization/scalability related aspects in RAN4 for AI/ML enabled CSI feedback enhancement, RAN4 should define different scenarios based on parameters listed in the tables above.   Less generalized functionalities across a set of scenarios can result in frequent switching of model/functionality resulting in performance degradation.   1. RAN4 needs to design a new metric, indicative of generalization capabilities of AI/ML model/functionality, to verify the generalization performance of the model/functionality in different scenarios.   **On LCM requirements:**  For UE-assisted or NW-based performance monitoring, if required LCM action is not taken in a timely manner, the performance of AI/ML-based CSI feedback may be degraded to undesirable levels.   1. Use-case specific core requirements should be considered to limit latency of LCM actions (e.g. activation, deactivation, fallback, switching etc.) for AI/ML-enabled CSI feedback. |
| [**R4-2412130**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412130.zip) | NTU | **Proposal 1: Study the feasibility of the partial specified test decoder with partial mapping alignment**  **For any latent message , where is the latent message space, denote the decoder output of the reference enc/dec pair as and the decoder output of the test decoder (from another pair) as , the test decoders has to satisfy: , where is the space of the test decoder output.**  **Observation 1: The advantages of partial mapping alignment approach for partial specified test decoder option are:**   * **A standardized training data free approach: 3GPP standard doesn’t have to capture the training data.** * **If the reference decoder can be described analytically with a set of equations, e.g., e-Type II CSI, RAN4 doesn’t have to agree a reference encoder/decoder pair to derive the test decoders.** * **Best forward compatible: either dataset or reference encoder/decoder pair approaches require UE vendors to train the decoder based on the encoder or dataset from a specific model designed based on the currently available AI/ML model structures. When new and superior models are developed by AI research communities, the dataset or reference encoder may limit the implementation flexibility of the UE vendors to adopt the new models.**   **Proposal 2: Additional margin to account for the mismatch need to be considered when determining requirement.**  **Proposal 3: RAN4 either agrees multiple options of test decoder model structures and other parameters, or agrees a reference decoder derivation procedure for work item phase, in the study phase to demonstrate the feasibility of CSI use case from the interoperability and verification perspectives.**  **Proposal 4: RAN4 uses the following steps, summarized from contribution [3,4,5,6] to derive the reference decoder (and encoder) in the WI phase and captures them in the TR. The open issues listed in each of the steps can be discussed in study or work item phase.**   * **Step 1: Determine simulation setup and encoder input dataset generation procedure**   + **Summary:**      - **Open issue for simulation setup:**  1. **System level or link level simulation for determining test decoder** 2. **Using setup in TR as starting points, what are the parameters need to be updated**    * + **Open issue for data collection/merge: whether RAN4 needs to collect and merge encoder input generated by different companies**  * **Step 2: Determine model type and parameter**   + **Summary of open issues:**  1. **Whether further update is needed for the table in R4-2403712** 2. **Whether multiple options can be listed as the outcome of this step, if RAN4 experiences difficulties to pick one before evaluation step: companies can pick their favorite models to come up with their proposals for evaluation in step 4**  * **Step 3: Determine training hyper parameters**   + **Summary of open issues:**  1. **whether RAN4 needs to agree training related parameters, and if needed, what parameters to be agreed** 2. **Whether RAN4 needs to agree on dataset including encoder output**   **(After step 3, companies can train and propose (encoder and) decoder)**   * **Step 4: Determine the test decoder(s) by first determine evaluation methodology and selection/merge criterion, and then decide the test decoder(s) based on companies’ evaluation**   + **Summary:**     - **Open issue on evaluation method and criterion: whether to consider one more factors listed in the following**  1. **Achievable performance metrics: SGCS or NMSE** 2. **Complexity: including flops or model storage size, can consider to set an upper bound** 3. **Robustness (based on performance metrics): whether the decoder can achieve satisfactory performance when connecting to encoder with different structures with the decoder, but trained with the decoder’s input and output dataset**    * + **Open issue on selecting and merging the decoder proposals: for each testing scenario, whether RAN4 directly selects one decoder proposal from all the proposals (after step 3) or RAN4 can merge parameters from multiple proposals with the same structure into on decoder based on further studies.**      + **Open issue on iterations: if RAN4 can’t agree or merge proposals to one decoder, whether RAN4 goes back to step 2/3 to agree more model/training parameters and repeat step 4.** |
| [**R4-2412252**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412252.zip) | vivo | **Proposal 1: Reference encoder and decoder are needed for requirement derivation, and could largely be used for test decoder derivation for both Option 3 and Option 4. RAN4 to work on reference encoder and decoder firstly. Later to discuss what will be put in the spec.**  **Proposal 2: Similar feasibility conclusion through the simulation campaign of test options can be draw for both Option 3 and Option 4.**  **Proposal 3: Reference encoder needs to be specified, at least to be used as verification encoder in the TE test decoder verification.**   * **No extra training procedure to train the verification encoder will be needed.** * **No performance degradation from imperfectly matched verification encoder will be caused.**   **Proposal 4: After the alignment of reference encoder and decoder, what will be put in the spec for Option 3 and Option 4 is listed in the following.**   * **Option 3: Test decoder (+ reference encoder model), and channel generation method are in the spec;** * **Option 4: Reference encoder (+ test decoder model structure), and channel generation method are in the spec.**   **Proposal 5: One test decoder could be used for one test case or multiple test case.**  **Proposal 6: Different reference encoder may be defined for different requirement or test case.**  **Proposal 7: As seen in Figure 2.2-1, the reference/test encoder/decoder may be aligned through the following procedures**   * **Step 1: Align the dataset containing only channel information.** * **Step 2: Determine the model hyperparameters that need to be aligned.** * **Step 3: Define the evaluation method for model complexity and performance.** * **Step 4: The best model structure(s) may be selected based on the aligned evaluation method, through the simulations using the aligned dataset.** * **Step 5: Based on the aligned model structure, the specific parameters of the reference model would be merged from companies**   **Proposal 8: Take into consideration the parameters to be aligned for Option 3 and Option 4 in Table 2.2-1.**  **Observation 1: Based on the above aligned model structure and simulation assumptions, the SGCS of AI/ML model for 64 bits payload is 0.75 and the SGCS of eType II is 0.73.**  **Proposal 9: In future actual reference encoder and reference decoder design, RAN4 may consider smaller model, for the convenience of implementation.**  **Proposal 10: Considering the number of RAN4 meetings and the number of companies, the storage size of dataset needed per year for RAN4 may be 100\*10GB= 1TB. The data format of training channel is suggested to be a matrix of M\*N, where M is the length of encoder input and N is the number of samples.**  **Proposal 11: Considering the number of RAN4 meetings and the number of companies, the storage size of model needed per year for RAN4 may be 100\*100MB= 10GB. ONNX is suggested as model format, since ONNX can be compatible with both PyTorch and TensorFlow.**  **Table 2.5-2. SGCS under different channel model for the proposed CNN and Transformer.**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Model structure** | | **CNN** | | **Transformer** | | | **CSI feedback bits** | **Training dataset** | **Test dataset** | | **Test dataset** | | | **CDL-C** | **UMa** | **CDL-C** | **UMa** | | **116 bits** | **CDL-C** | **0.994** | **0.336** | **0.996** | **0.253** | | **UMa** | **0.791** | **0.787** | **0.831** | **0.821** | | **Mixed** | **0.982** | **0.773** | **0.988** | **0.801** | | **CNN encoder: FLOPS 126M, size 0.1M. CNN decoder: FLOPS 1974M, size 2.4M.**  **Transformer encoder: FLOPS139M, size 5.6M. Transformer decoder: FLOPS 139M, size 5.6M**  **Note: Mixed training dataset includes CDL-C and UMa channel model.** | | | | | |   **Proposal 12: Using the mixed dataset for model training, including the mixing of TDL, CDL and UMa, while using the TDL dataset for RAN4 tests. Other mixing rules are not precluded.**  **Observation 2: From initial results for field test, the generalization performance of AI/ML model trained by UMa simulation data on field data seems acceptable, which has similar performance as eType II codebook. The generalization performance of AI/ML model trained by CDL simulation data on field data is worse than AI/ML model trained by UMa simulation data.**  **Proposal 13: “Supported training collaboration type between DUT and decoder provider” can be removed from the table of the comparison of the four options of test decoder, since this aspect is just for training before test and seems to have no obvious impact on the test.**  **Proposal 14: Take into consideration the summary of 4 options for testing of 2-sided model in Table 2.7-1.**  **Proposal 15: Compared with absolute throughput, relative throughput would be used to see the gain from CSI prediction. The comparison baseline can be further discussed, such as randomly chosen PMI.**  **Proposal 16: Since monitoring is still under discussion in RAN1, RAN4 should wait for further progress of RAN1.** |
| [**R4-2412332**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412332.zip) | Intel Corporation | **Proposal #1: Further align the AI/ML test methodology with RAN1's conclusions on inter-vendor training collaboration for AI/ML-based CSI compression.**  **Proposal #2: Discuss and agree on the detailed procedure for Option 3 feasibility analysis.**   * **Agree on the simulation assumptions for dataset generation (completed)** * **Agree on the AI-ML model assumptions for feasibility analysis (completed)** * **Agree on the procedure for vendor-specific encoder training** * **Agree on the framework/format for AI-ML model sharing among companies** * **Agree on the framework for dataset sharing among companies including specific set of datasets to be used for encoder training** * **Select one or more decoder models for further analysis**   + **Companies bring simulation results for the pair of trained Encoder/Decoder models**     - **Check and confirm simulation results alignment**     - **If alignment is not achieved, agree on modified/additional simulation assumptions and/or model assumptions for the next round.**   + **Select one or more decoder models for further analysis**   + **Agreed models are shared with other companies based on the agreed framework/format** * **Collect datasets for model training** * **Check feasibility of Option 3 (i.e. to design/train UE-specific encoder based on the agreed common Decoder)**   + **Agree on the criteria to confirm feasibility of Option 3**   + **Companies bring simulation results for the pair of trained company-specific Encoder and common Decoder models**   + **Check feasibility based on the agreed criteria**   **Proposal #3: Update the simulation assumptions for Option 3 feasibility analysis based on Table 1.**  Model complexity analysis   |  |  |  | | --- | --- | --- | |  | **Number of trainable parameters** | **MFLOPs** | | **Encoder** | 125 793 | 46.62 | | **Decoder** | 1 804 482 | 738.58 |   Simulation results:   |  |  |  | | --- | --- | --- | | **Scenario** | **Mean SGCS** | | | **AI CSI feedback (RAN4 model)** | **eTypeII CSI feedback** | | 2GHz / 15kHz SCS / 10MHz BW / layer 1 | 0.689 | 0.658 | | 2GHz / 15kHz SCS / 10MHz BW / layer 2 | 0.517 | 0.512 |   **Proposal #4: Discuss and agree on the detailed training procedure for vendor-specific encoder.**  **Proposal #5: Further investigate and select the approach to share AI-ML models for CSI Decoder including**   * + **Option 1: Share model parameters in PyTorch or TensorFlow formats and source codes**   + **Option 2: Share entire models in PyTorch or TensorFlow formats (no source codes)**   + **Option 3: Share models in ONNX format (no source codes)**   **Proposal #6: Further discuss and agree on the format for dataset sharing/alignment for Option 3 feasibility analysis.**  **Proposal #7: Further discuss and define the upper bound complexity including the number of computations and number of parameters for test/reference encoders/decoders for the purpose of performance requirements definition.** |
| [**R4-2412609**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412609.zip) | Ericsson | [Observation 1 The test decoder will set the latent space that is contained within real encoders and decoders.](#_Toc174116842)  [Observation 2 The test decoder and the latent space it represents must be sufficiently rich to be useful for real world operation.](#_Toc174116843)  [Observation 3 A large amount of discussion will be needed to fix assumptions on model structure, complexity and some detailed implementation aspects in a real WI.](#_Toc174116844)  [Observation 4 The test decoder need not be interoperable with all reference encoders considered during the RAN4 evaluation.](#_Toc174116845)  [Observation 5 If convergence is observed in performance of different encoder/decoder pairs under the same evaluation conditions then the difference between different companies encoder/Decoder models is in the latent space representation, but not the performance.](#_Toc174116846)  [Observation 6 Once the standardized test decoder is available, encoder vendors can create real encoders that can interoperate with the standardized encoder.](#_Toc174116847)  [Observation 7 There is a need to carefully consider a representation for datasets that enables wide and open access.](#_Toc174116848)  [Observation 8 There is a need to consider a model representation that enables a wide and open access.](#_Toc174116849)  [Observation 9 The framework used for sharing models may depend on the purpose of sharing (performance verification, detailed implementation in hardware etc.)](#_Toc174116850)  [Observation 10 A common graph representation of the shared ML model might be used to describe completely the models architecture.](#_Toc174116851)  Based on the discussion in the previous sections we propose the following:  [Proposal 1 For option 3, capture both the test decoder and associated reference encoder in the TS.](#_Toc174116852)  [Proposal 2 Consider the simulation assumptions and criteria for selection of the test decoder in RAN4 separately to the simulation assumptions and criteria for setting the performance requirement.](#_Toc174116853)  [Proposal 3 Base the simulations for evaluation of which test decoder to select on system data in order that the selected test encoder can be shown to have reasonable performance across the range of operating conditions.](#_Toc174116854)  [Proposal 4 For the performance requirement, discuss to use system data and BLER as a metric.](#_Toc174116855)  [Proposal 5 If link level evaluation is needed for performance requirements, define link level assumptions for a wide spread of channel conditions, Doppler, SINR etc.](#_Toc174116856)  [Proposal 6 For option 4, capture the reference encoder in the TS.](#_Toc174116857)  [Proposal 7 Use the same procedure for identifying an encoder/decoder pair in RAN4 for either option 3 or option 4.](#_Toc174116858)  [Proposal 8 An activation function is agreed before the uniform scalar quantization](#_Toc174116859)  [Proposal 9 A dataset size is identified](#_Toc174116860)  **Simulation parameters and results:**  **Table 2. Hyper-parameters used during the training.**   |  |  | | --- | --- | | **Hyper parameters** | **Values** | | Batch size | 32 | | Loss function | NMSE | | Optimizer | Adam | | Learning rate | Warmup cosine decay, 10-4 to 10-5 |   **Table 3. Evaluation results for the agreed CNN model**   |  |  |  |  | | --- | --- | --- | --- | | **Compression method** | **# Training dataset** | **SGCS** | **Gain** | | Rel. 16 eType II PC 1 | #N/A | 0.7001 |  | | AI-based compression (64 bit) | 200 K | 0.6704 | -4.2% | | 400 K | 0.6891 | -1.6% | | 800 K | 0.7085 | 1.2% |   **Table 3. Evaluation results for the agreed CNN model with beam-delay domain inputs**   |  |  |  |  | | --- | --- | --- | --- | | **Compression method** | **# Training dataset** | **SGCS** | **Gain** | | Rel. 16 eType II PC 1 | #N/A | 0.7001 |  | | AI-based compression (64 bit)  - W2-like input | 100 K | 0.7178 | 2.5% | | 200 K | 0.7235 | 3.3% | |
| [**R4-2412768**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412768.zip) | Huawei,HiSilicon | ***Observation 1:* MLP is superior to CNN in terms of SGCS.**  **Table 1: SGCS performance of MLP, CNN and Transformer**   |  |  |  |  | | --- | --- | --- | --- | |  | **MLP** | **CNN** | **Transformer** | | **AI** | **0.70** | **0.69** | **0.70** | | **R16 Type-II** | **0.66** | | |   ***Observation 2:* The complexity of MLP is lower than that of CNN.**  ***Proposal 1:* Principle of reference model design – complexity of reference model is one of the metrics for selecting reference model structure.**  ***Proposal 2:* Different model structures of encoder and decoder can be considered for reference model design, e.g., MLP for encoder while CNN for decoder.**  ***Proposal 3:* Take Option 3a as baseline, where a specific rather than a mixed dataset is used for defining the test decoder in each test case.**  ***Proposal 4*: For achieving a converged test decoder in Option 3, at least the structure of both the reference encoder and test decoder, hyperparameters of model training, as well as a determined sample-by-sample dataset are expected to be aligned among companies.**  ***Observation 3*: Even with all hyperparameters aligned and model training converged, the model parameters provided by companies can still be different. How to align model parameters of the test decoder among companies is an open issue.**  ***Proposal 5*: Reuse RAN4 legacy PMI reporting test configuration in 38.101-4.**  ***Observation 4*: Different from RAN1 purpose, the goal in RAN4 is to define the minimum requirement of throughput for PMI reporting, where the TDL is sufficient as proved in legacy.**  ***Proposal 6*: Take relative throughput compared to eType-II as the test metric for AI-CSI.**  ***Observation 5*: If using CDL, it is challenging to achieve the aligned throughput performance of eType-II for deriving the relative throughput of AI-CSI.**  ***Proposal 7*: Avoid using CDL for deriving the performance requirement in AI-CSI, where the baseline performance (i.e., throughput of eType II under CDL) is missing.**  ***Proposal 8*: According to whether the model structure is specified, Option 4 can be further divided into two sub options as follows.**   * **Option 4a: Model structure is not specified in RAN4. Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder.** * **Option 4b: Model structure is specified in RAN4. Training dataset is not specified for verifying the encoder at DUT. The test decoder developed by TE vendor needs verification.**    + **FFS: How to determine the test metric for test decoder developed by each TE vendor.**   ***Observation 6*: The boundary between Option 3 and Option 4b is whether the model parameters are specified in RAN4.**  ***Proposal 9*: In Option 4, the performance of test decoder should be verified before testing DUT in each test.**  ***Proposal 10*: Compression ratio and quantization level needs to be specified in Options 3 and 4.**  ***Observation 7*: Though there is no need to align the model parameters of test decoder, model parameters of the reference encoder for verifying the test decoder is still needed. How to align is still an open issue.**  ***Proposal 11*: Deprioritize SCGS/NMSE for defining baseline requirements in AIML-enabled CSI compression.**  ***Proposal 12*: Deprioritize CSI prediction accuracy for defining baseline requirements in AIML-enabled CSI prediction.**  ***Observation 8*: How to ensure that the testing dataset aligns well with training dataset is still an open issue.** |
| [**R4-2412792**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2412792.zip) | ZTE Corporation, Sanechips | ***Observation 1. AI CSI compression indeed achieves performance improvement compared to legacy eType II codebook from simulation results.***  ***Observation 2. RAN4 built a strict guidelines to govern process for Release 15 for demodulation requirements.***  ***Observation 3. If RAN4 would like to study feasibility for CSI compression, new regulations should be considered for AI-based KPI.***  ***Observation 4. RAN4 requires more evaluations to validate the impacts of such differences on the outcomes.***  ***Proposal 4. To consider standardizing the reference model or reference encoder and latent space for option 4.***  **Table 1. Simulation results for rank1**   |  |  |  |  | | --- | --- | --- | --- | | **Model structure** | | **CNN** | **eType II** | | **CSI feedback bits** | **Training dataset** | **Test dataset** | **/** | | **UMa** | **UMa** | | **64 bits** | **UMa** | **0.7294** | **0.6803** | | **FLOPS and size** | | **FLOPS:**  **Encoder: 46.966M**  **Decoder: 739.965M**  **Parameters Num: 1.928M** | **/** |     **Table 2. Simulation results for rank 2**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Model structure** | | **CNN** | | **eType II (118bits)** | | | **CSI feedback bits** | **Training dataset** | **Test dataset** | | **/** | | | **UMa** | | **UMa** | | | **64 bits+64bits** | **UMa** | **Rank 1** | **0.7228** | **Rank 1** | **0.6870** | | **Rank 2** | **0.5865** | **Rank 2** | **0.5538** | | **FLOPS and size** | | **FLOPS:**  **Encoder: 46.966M**  **Decoder: 739.965M**  **Parameters Num: 1.928M** | | **/** | | |
| [**R4-2413169**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112/Docs/R4-2413169.zip) | Qualcomm Incorporated | **Observation 1: Following training parameters were used to generate results for the CNN based encoder and decoder pair with the agreed model.**   |  |  | | --- | --- | | **Parameter** | **Value/Name** | | **Training algorithm** | **Adam optimizer** | | **Learning rate of training algorithm** | **10^-4** | | **Number of epochs** | **150** | | **Batch size** | **128** | | **Training set** | **450k** | | **Training loss function** | **SGCS + additional terms for quantization aware training** |   **Observation 2: SGCS results of AI/ML encoder-decoder and Rel-16 eType2 CSI are as follows:**   |  |  |  | | --- | --- | --- | | **Model** | **SGCS (layer 0)** | | | **CNN encoder – CNN decoder**  **(generated based on the agreements of the last meeting)** | **0.6934** | **-5.13 dB** | | **Rel-16 eType2** | **0.6701** | **-4.82 dB** | |

## Open issues summary

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

1. Standardization steps for Option 3
2. Simulation results discussion
3. Criteria to select decoder for feasibility study
4. Option 4 feasibility study
5. CSI prediction accuracy metric
6. Test decoder verification
7. Generalization issues

### Sub-topic 4-1

*Option 3 feasibility study steps*

In the previous meeting it was agreed to perform a feasibility study for Option 3(fully specified decoder) based on “Option 1” as follows:

Option 1

companies bring encoder + decoder set based on agreed parameters. RAN4 chooses one of the decoders and interested companies further check if an encoder can be trained with this decoder to obtain similar performance/complexity (or other evaluation criteria)

The next steps should be discussed and agreed

**Issue 4-1: Option 3 next steps**

* Proposals
  + Option 1:
* **Agree on the simulation assumptions for dataset generation (completed)**
* **Agree on the AI-ML model assumptions for feasibility analysis (completed)**
* **Agree on the procedure for vendor-specific encoder training**
* **Agree on the framework/format for AI-ML model sharing among companies**
* **Agree on the framework for dataset sharing among companies including specific set of datasets to be used for encoder training**
* **Select one or more decoder models for further analysis**
  + **Companies bring simulation results for the pair of trained Encoder/Decoder models**
    - **Check and confirm simulation results alignment**
    - **If alignment is not achieved, agree on modified/additional simulation assumptions and/or model assumptions for the next round.**
  + **Select one or more decoder models for further analysis**
  + **Agreed models are shared with other companies based on the agreed framework/format**
* **Collect datasets for model training**
* **Check feasibility of Option 3 (i.e. to design/train UE-specific encoder based on the agreed common Decoder)**
  + **Agree on the criteria to confirm feasibility of Option 3**
  + **Companies bring simulation results for the pair of trained company-specific Encoder and common Decoder models**
  + **Check feasibility based on the agreed criteria**
  + Option 2:

A screenshot of a computer

Description automatically generated

* + Option 3: others
* Recommended WF
  + Discuss the next steps:

1. Check on performance alignment -> see simulation results from contributing companies
2. Share decoders/datasets (training and testing) – logistics to be discussed separately
3. Select one or more decoder for further analysis
   * selection criteria to be discussed
   * decoder to be shared with other companies
4. Each company brings results for training of “own encoder” with selected decoder
   * performance alignment to be checked/discussed
5. Check overall feasibility of Option 3

### Sub-topic 4-2

*Simulation results discussion*

Several companies brought simulation results based on the agreed parameters in the email discussion and shown in R4-2411786.

**Issue 4-2: Simulation results**

* Proposals
  + Option 1: Check on the simulation results
* Recommended WF
  + To be discussed whether alignment is good enough or not

### Sub-topic 4-3

*Criteria to select decoder for feasibility study*

One of the steps of the Option 3 feasibility study will be to select a decoder for companies to train an encoder. Selection criteria should be discussed. This decoder and the criteria is only to be used for this feasibility study, the selection criteria will be rediscussed in a potential WI as needed.

**Issue 4-3: Decoder selection criteria**

* Proposals
  + Option 1: Best performance (SGCS) and mean performance of corresponding eType II CSI feedback
  + Option 2: select multiple and work in parallel
    - few decoders can be selected and a subset of companies check feasibility on one decoder
  + Option 3: random selection, as long as performance is similar it does not matter which decoder is selected
  + Option 4: other criteria
* Recommended WF
  + To be discussed

### Sub-topic 4-4

*Options for Option 4*

**Issue 4-4: Options for Option 4**

* Proposals
  + Option 1:
  + Option 2:
  + Option 3:
  + Option 4:
* Recommended WF
  + To be discussed, options are not exclusive

### Sub-topic 4-5

*CSI prediction accuracy metric*

**Issue 4-5: CSI prediction accuracy metric**

* Proposals
  + Option 1:
  + Option 2:
  + Option 3:
  + Option 4:
* Recommended WF
  + To be discussed, options are not exclusive

### Sub-topic 4-6

*Test decoder verification*

**Issue 4-6: Test decoder verification**

* Proposals
  + Option 1:
  + Option 2:
  + Option 3:
  + Option 4:
* Recommended WF
  + To be discussed, options are not exclusive

### Sub-topic 4-7

*Generalization issues*

**Issue 4-7: Generalization Issues**

* Proposals
  + Option 1:
  + Option 2:
  + Option 3:
  + Option 4:
* Recommended WF
  + To be discussed, options are not exclusive