**3GPP TSG-RAN WG4 Meeting #112bis R4-241xxxx**

**Hefei, Anhui, China, 14th – 18th October, 2024**

**Agenda item:** 6.17.6

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [112-Bis][127] 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 R4-2414323. This summary is organized in 3 high level topics and contains several sub-topics for discussion.

# Topic #1: General aspects and TR

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2414958**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2414958.zip) | CAICT | **Proposal 1: Consider effective performance monitoring for post-deployment handling. The effectiveness of performance monitoring is guaranteed by specifying requirements and tests for monitoring and LCM procedure.**  **Proposal 2: Suggest to discuss potential solutions to build dataset for offline performance tests conducted inside vendor.**  **Proposal 3: Fallback to non-AI mode can be taken as baseline mechanism for post-deployment handling.**  **Proposal 4: For generalization tests, “other scenarios and/or configurations” are the scenarios and/or configurations that are not reported as “applicable”. Further discuss whether to include scenarios and/or configurations that are not reported as “supported” on UE side.**  **Proposal 5: For generalization tests, legacy non-AI performance requirement could be considered for “other scenarios and/or configurations”.** |
| [**R4-2415169**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415169.zip) | CATT | **Observation 1: The number of test configurations/parameters depends on the granularity of functionality/ feature, pending on RAN1 discussions.**  **Proposal 1: RAN4 discusses the number of test configurations/parameters per use case after RAN1 agrees on the granularity of AI/ML functionality/feature.**  **Proposal 2: Better performance can be expected for positioning and CSI prediction use case. For BM, UE is required to fulfil the legacy performance requirements. FFS for CSI compression in WI, if there will be one.**  **Proposal 3: RAN4 to deprioritize the discussions on generalization tests until RAN4 completes performance requirements and figures out how to define the tests without generalization.**  **Proposal 4: Postpone LCM discussions until more RAN2/1 agreements are made. When LCM discussions start, RAN4 can focus on one use case and extends the LCM requirements to other use cases, if necessary.** |
| [**R4-2415186**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415186.zip) | vivo | **Proposal 1: For verifying the minimum performance of AI/ML functionality/feature, RAN4 to first determine the factors to be considered in the test design. For each use case, the candidate configurations/scenarios include:**   * **For beam management**   + **Deployment scenarios: different ISD, UMi/UMa**   + **Configurations:**     - **various gNB settings:**        * **gNB antenna array dimensions**       * **DL Tx beam codebook (e.g., beam width, beam angle, etc)**     - **various Set B patterns**     - **various Set A patterns** * **For CSI compression**   + **Channel model: InH, UMi, UMa, TDL and CDL**   + **Configurations:**     - **various TxRU mappings**     - **various CSI payload sizes**     - **bandwidths**     - **Tx port numbers** * **For CSI prediction**   + **Channel model: UMi, UMa, TDL and CDL**   + **Configurations:**     - **various UE speed**     - **various carrier frequency** * **For positioning**   + **Case 1:**     - **Channel model:**       * **Drop**       * **Scenarios: SH, DH, HH, etc**       * **LOS/NLOS**       * **Delay spread**     - **TRP deployment:**       * **TRP number**       * **TRP location**   + **Case 2:**     - **Channel model:**       * **NLOS channel or AWGN**   + **Case 3:**     - **No need to test** |
| [**R4-2415253**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415253.zip) | CMCC | ***Observation 1: only rely on conformance testing cannot solve the post-deployment handling issue, since the test burden is huge and compliance testing can not test model performance in all configurations/scenarios.***  ***Observation 2: only rely on LCM cannot solve the post-deployment handling issue, since LCM target to monitor whether there is performance drift for serving model/functionality, while the motivation to have post deployment handling is to check the validity (e.g.minimum performance) of updated/new model before it is loaded to the UE deployed in the field.***  ***Proposal 1: for post deployment handling, the solution to ensure that a validated model exists at the UE and/or the solution to ensure that non-validated/non-tested models are not used by the UE can be considered.***  ***Proposal 2: it is proposed that different test configurations/parameters are used for tests.***  ***Proposal 3: The issue on how many different test configurations/parameters are used for tests can be further discussed based on RAN1/2 progress.***  ***Proposal 4: for generalization, it is proposed to take the requirements for inference as the minimum level performance for generlazation.*** |
| [**R4-2415271**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415271.zip) | Korea Testing Laboratory | **Observation 1: There are numerous combinations of NW-side additional conditions and UE-side additional conditions.**  **Observation 2: The additional conditions must be defined at least for the conformance testing when the associated ID is used between a test setup and a DUT.**  ***Proposal 1: It is necessary for RAN4 to study and discuss the NW-side additional conditions and the UE-side additional conditions for the conformance testing.***  **Observation 3: The AI/ML performance provides the stochastic performance variation.**  **Observation 4: It has different behavior to verify the minimum performance of AI/ML functionality/feature** **depending on “Signalling procedures for model and functionality LCM” in TR 38.843.**  ***Proposal 2: It should be discussed within RAN4 group for the way of efficiency conformance test using the mechanism of the fallback / the deactivation if it might not be possible to define the mechanism in detail since the reason of the implementation area.***  ***Proposal 3: It is efficiency way for the conformance test such as the performance test, except the protocol test, to decide one of between “Network decision, network-initiated AI/ML management (replace Network to TE)” and “UE decision, event-triggered as configured by the network (replace Network to TE)”***  **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 can effectively prevent overfitting. If there is uncertainty, other approaches should be taken into account.***  ***Proposal 5: It is necessary to validate whether a DUT has the overfitting problem when test configurations may not completely prevent overfitting.*** |
| [**R4-2415495**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415495.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**  ***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: For the purpose of limiting the amount of data needed for fine-tuning and reducing the number of fine-tuning iterations investigate transfer learning and meta-learning for efficient re-learning across different scenarios to aid generalization**  **Proposal 16: 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 17: 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 18: 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 19: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case**  **Proposal 20: 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** *Post Deployment validation and On-Device Fine tuning*  **We are making a new proposal for post deployment (Option 1d) combined with on-device fine tuning. Based on the principles below:**  **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.**  ***Proposal 26: (Option 1d)* For the post deployment framework, investigate an on-device model monitoring and performance assessment framework to track both active and inactive AI/ML models that experience actual UE hardware and field data. The aim is to dynamically manage the model database, enabling proactive monitoring for computing and updating key performance indicators (KPIs). This will facilitate seamless transitions to newly updated and fine-tuned models as needed.**  **To minimize the number of samples required and reduce training costs, the framework will leverage Transfer Learning and Meta Learning techniques, ensuring efficient model management. The ultimate goal is to optimize performance, reduce overhead, and ensure models adapt seamlessly to changing conditions for post deployment and fine-tuning.**   * + **KPI can be signaled to the NW, KPIs are used to update the monitoring scores for each model (updates the database of active and inactive models)**   + **UE signals the monitoring scores to NW and based on those KPIs, the NW can signal the new model ID to be deployed** **or updated/fine-tuned** or signal the UE or switch back to legacy operation.  (through RRC signaling)   + **NW can predict performance degradation (in future) and signal UE to a) switch model or (b) to identify which model to fine-tune**   **Proposal 27 *(Option 1d)* : 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 company  Description automatically generated Proposal 28 *(Option 1d)*: In the post-deployment framework, models would be allowed to undergo fine-tuning, utilizing the model monitoring dataset for ongoing fine-tuning and re-training purposes. **Proposal 29: In the post-deployment framework, the network (NW) can indicate which model should be fine-tuned, with the fine-tuning process occurring either at the UE or within the network itself or an OTA server**  **Proposal 30 *(Option 1d)*: If fine-tuning takes place at the network (NW) or an over-the-air (OTA) server, the UE transmits a small sample dataset with high precision (e.g., PMI) over the air. The NW or OTA server (where inference takes place) then performs fine-tuning using transfer learning or meta learning techniques on the identified model to be fine-tuned, minimizing the number of parameters that need updating. It subsequently requests another dataset for validation to validate the model for post deployment. The updated set of parameters is then downloaded to the UE, with the option to quantize the coefficients to reduce overhead.**  **Proposal 31 *(Option 1d)*: If fine-tuning occurs at the UE, the UE is responsible for performing the fine-tuning using a small sample training dataset, with the network (NW) signaling which model to undergo updating. The UE then performs model monitoring to verify and validate the model post-deployment. The goal is to achieve optimal model performance as quickly as possible, minimizing the number of samples required during Transfer Learning or Meta Learning to reduce training costs and achieve fast adaptation.** *2.7 Data collection for testing* **Proposal 32: 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.** |
| [**R4-2415620**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415620.zip) | Huawei,HiSilicon | According to the discussion, following proposals and observations are provided:  ***Proposal 1:*** If the AI related configurations/parameters is a subset of (or the same as) the legacy test configurations/parameters, take the existing test configurations/parameters as a starting point.  ***Proposal 2:*** If there exists legacy requirement, AI requirement has to be higher than that of non-AI.  ***Observation 1:*** For some cases, to verify AI performance, new test set up is under discussion.  ***Proposal 3:*** 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. 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 verify the AI benefits. |
| [**R4-2415687**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415687.zip) | Intel Corporation | **Proposal #1: Coordinate with RAN1/2 on the following questions related to post-deployment model performance monitoring and LCM procedures:**   * + **Recommend RAN1/2 to ensure that the performance monitoring framework for AI/ML models deployed at the UE can be used to compare the performance of AI-ML model and non-AI-ML performance for use cases where legacy behavior is defined**   + **Check whether it is feasible that UE supports at least some default/baseline AI/ML models/features that passed conformance requirements, and the respective models can be used as a fallback mechanism for certain scenarios**   **Proposal #2: RAN4 to confirm that changes or updates to the AI/ML models/features can be tested by the device vendors against requirements before the deployment to the UE is performed**   * + **The conditions when testing can be performed are FFS**   + **The information on whether AI/ML model/feature update has passed conformance test (and potentially associated data) shall be conveyed to the network.**   **Proposal #3: 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 for all use cases.** |
| [**R4-2415930**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415930.zip) | NTT DOCOMO, INC. | **Observation 1: It is difficult to cover all new/updated AI model/functionality by conformance test before deployment from both testing effort and consuming time points.**  **Proposal 1: If the performance degradation level is critical, the AI model/functionality should be deactivated or fallback to the default/initial AI model/functionality during performance monitoring. The monitoring accuracy and fallback execution delay are needed to be specified.**  **Proposal 2: For reactivating the deactivated AI model/functionality, performance monitoring is needed to be done in parallel with normal measurement, i.e. if the inference accuracy is good enough compared with exact measured value, the AI model/functionality is reactivated.**  **Proposal 3: If a usecase employs legacy metric, the AI/ML model performance should be at least as same as that of legacy. If a usecase employs new metric, the comparison between legacy non-AI/ML requirements is not needed.**  **Proposal 4: Generalization test for static scenario is not necessarily to be specified. As same principle as legacy test, it does not mean the acceptance of significant performance degradation in the condition other than the specified test condition.** |
| [**R4-2415964**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415964.zip) | Nokia | The following Observations and Proposals were made:  **On functionality-based conformance testing:**  **Proposal 1:** **RAN4 to confirm that the new agreement on functionality-based testing goal is substituting Option 1 and Option 2 testing goals considered in Rel-18 and listed in the TR 38.843.**  **Observation 1**: The way how requirements in RAN4 will be defined will depend on the way how Associated IDs will be standardized, e.g., whether any corroding parameters related to Associated ID will be specified or will they be fully vendor specific.  **Proposal 2: Each specified AI/ML-enabled functionality shall have RAN4 requirements and pass at least one test for device conformance for this functionality.**  **Proposal 3: A minimum set of Associated IDs should be defined as mandatory (mandatorily applicable) for testing of each functionality.**  **Observation 2**: A considerable change of conditions that require the change of Associated ID may require RRC reconfiguration.  **Proposal 4: RAN4 does not need to define performance requirements that imply the change of Associated ID during the test.**  **Observation 3**: The number requirements per given UE AI/ML-enabled functionality and Associated ID will depend on their concrete definition in the other WGs and the generalization capabilities of the UEs that require further study in per use-case manner.  **Proposal 5: RAN4 will need to wait for more progress in specification of concrete AIML-enabled features and Associated IDs and study UE generalization capabilities to conclude about the number of requirements/tests per given functionality and Associated ID.**  **Proposal 6: Non-static radio scenarios/conditions for a given UE AI/ML-enabled functionality and Associated ID should be discussed separately from the study of AI/ML-enabled functionality generalization to different but static conditions (legacy way of testing).**  **On post-deployment testing:**  **Proposal 7: Option 1 of post-deployment testing shall be applicable only to changes in UE AI/ML-enabled functionality following the agreed testing goal definition, i.e., Conduct the testing for changed AI/ML functionality before its deployment in already deployed UEs (considering the UE hardware that it will be used in).**  **Observation 4:** A need for post-deployment testing will depend on the reason/type of change in the functionality or in its applicability or whether the change has been already tested or not (e.g., offline).  **Proposal 8: In Option 2, RAN4 to consider a cause/reason of the change in AIML functionality when deciding about a need for post-deployment testing.**  **Observation 5**: Option 1 of post-deployment discussion can be feasible only for the major implementation updates of the UE AI/ML-enabled UE functionalities.  **Observation 6**: Option 2 based on monitoring does not completely address the issue when changed UE functionality is already used in the live network without testing and could demonstrate performance below minimal requirements.  **Proposal 9: RAN4 to ensure that at least one functionality (for an AI/ML-enabled Feature) that has passed conformance testing shall be present in the device.**  **Proposal 10: RAN4 to consider handling post-deployment verification/testing of updated/new UE AI/ML functionalities in inactive/standby state/mode in parallel with active functionalities.**  **Proposal 11: 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.** |
| [**R4-2416004**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416004.zip) | Samsung | **Observation 1: RAN4 discussion on the post deployment testing has just been focused on the necessity of the consideration rather than the feasibility of the validation, and the feasibility has not been provided yet.**  **Observation 2: RAN4 could not reach a consensus on the post deployment testing due to lack of feasibility and other WGs’ progress.**  **Observation 3: Option 1 of post-deployment handling will incur huge testing burden if the conformance testing shall be conducted in certificated labs for every AI/ML model update/transfer.**  **Observation 4: It would be meaningless for RAN4 to discuss the test for monitoring performance or monitoring based solutions without monitoring and LCM discussion in other WG, which would be for the same issue in the end.**  **Proposal 1: Option 1 of post-deployment handling is preferred only if**  **- The conformance testing includes the validation testing performed by UE vendors in the development phase of AI/ML model, rather than formal conformance tests in certificated labs.**  **Proposal 2: Otherwise, RAN4 should pause the post deployment discussion until the progress in other WG while keeping the previous tentative agreement with Option 1 and Option 2.**  **Proposal 3: The latest agreements for general aspects, e.g., testing goal and channel modelling, shall be kept as they are until further progress can be achieved in RAN4 and/or other WGs.** |
| [**R4-2416202**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416202.zip) | Ericsson | **Issue 1-1: Post deployment testing** (tentative agreement) [1]   * **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 * **Option 3**: Capture model input during conformance testing for later testing of new models. * Other options are not precluded. * ***Proposal 1****: With Option 1, it is challenging to ensure performance in all scenarios not known a priori. The tested scenarios are likely to be limited to the ones defined for RAN4 requirements. This option also implies a greater amount of logistics for updating and deploying new models.* * ***Proposal 2****: Option 2 provides means to verify the model’s ability to perform KPI-based performance monitoring and its ability to act in a certain way upon KPI degradation. However, it is also important to ensure and verify the reliability of performance monitoring and the accuracy of reports used for KPIs,* for which RAN4 and RAN1 may need to work together to derive KPIs, measurements, and corresponding requirements*.*   + *FFS: KPIs that isolate the model behaviour from other factors.*   + *FFS: ensuring smooth fallback/transition between AI and non-AI model.* * ***Proposal 3****: Option 3 provides means for a fair comparison between different models in the same UE or similar* (e.g., from the architecture and RF/baseband perspective) *UEs. This option is advantageous when model comparison is needed, e.g., to ensure that a newly deployed model is not worse than the previous model. The principles of Option 3 can be used also together with Option 1 and Option 2.* * *Proposal 4: RAN4 sends an LS to inform other groups (e.g., RAN1 and RAN2) about its observations on feasibility, complexity, advantages, and disadvantages of Option 1, Option 2, and Option 3, to facilitate their discussions and to avoid standardizing solutions which are not feasible from RAN4 perspective.* * *Proposal 5: RAN4 to study the performance expectations and potential requirements for the following transitions:*   + *AI/ML mode → non-AI/ML mode,*   + *Non-AI/ML mode → AI/ML mode.* * *Proposal 6: RAN4 will study the desired performance gap between AI/ML and non-AI/ML operation modes and how it can be evaluated and achieved (which may require the same KPI(s) applicable for both AI/ML and non-AI/ML modes). This would be needed at least for UE-triggered fallback.* * *Proposal 7: 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 8: 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.* |
| [**R4-2416366**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416366.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.** |

## Open issues summary

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

1. LCM handling
2. Post deployment testing
3. LCM requirements
4. AI/ML performance
5. AI/ML test setup
6. Testing goals

### Sub-topic 1-1

LCM Handling and coordination with other groups

Some companies proposed to coordinate with other working groups regarding LCM

**Issue 1-1: LCM Handling and coordination with other groups**

* Proposals
  + Option 1: RAN4 to send an LS to RAN1/2 to recommend them to ensure that monitoring framework enables the UE to compare performance with non-AI/ML performance
  + Option 2: RAN4 to send an LS to RAN1/2 to ask them to consider conformance testability when designing LCM operations
  + Option 3: no need to send any LS, LCM definition is not in RAN4 competence
  + Option 4: others
* Recommended WF
  + To be discussed

### Sub-topic 1-2

*Post deployment testing*

Post deployment testing has been discussed for several meetings, and some options were agreed to be further considered, however, no concrete feasibility analysis was brought up yet. In this meeting, some other options (option 1d) and an LS to other groups are further proposed.

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

* Proposals
  + Option 1: Include “Option 1d” in R4-2415495 in a further study
* ***Proposal 26: (Option 1d)* For the post deployment framework, investigate an on-device model monitoring and performance assessment framework to track both active and inactive AI/ML models that experience actual UE hardware and field data. The aim is to dynamically manage the model database, enabling proactive monitoring for computing and updating key performance indicators (KPIs). This will facilitate seamless transitions to newly updated and fine-tuned models as needed.**
* Option 2: *RAN4 sends an LS to inform other groups (e.g., RAN1 and RAN2) about its observations on feasibility, complexity, advantages, and disadvantages of Option 1, Option 2, and Option 3, to facilitate their discussions and to avoid standardizing solutions which are not feasible from RAN4 perspective.*
  + Option 3: Postpone this discussion until some more concrete proposals are made on how any of the options would work in practice
* Recommended WF
  + Option 3

### Sub-topic 1-3

*LCM requirements*

Several companies proposed to study various potential requirements while there are also proposals to postpone the discussion until more details are known about the framework defined by other groups

**Issue 1-3: Potential LCM requirements**

* Proposals
  + Option 1: Study requirements for reactivated a deactivated AI model functionality
    - **performance monitoring is needed to be done in parallel with normal measurement, i.e. if the inference accuracy is good enough compared with exact measured value, the AI model/functionality is reactivated.**
  + Option 2: 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).
  + Option 3: Postpone LCM discussions until more RAN2/1 agreements are made. When LCM discussions start, RAN4 can focus on one use case and extends the LCM requirements to other use cases, if necessary.
  + Option 4: others
* Recommended WF
  + Option 3

### Sub-topic 1-4

*Considerations on AI/ML performance*

**Issue 1-4: AI/ML Performance**

* Proposals
  + Option 1: AI/ML based performance should be better than the legacy (non AI/ML based) if an equivalent legacy requirement exists
  + Option 2: others
* Recommended WF

Take option 2 as baseline

### Sub-topic 1-5

*AI/ML test setup and performance*

**Issue 1-5: AI/ML test setup**

* Proposals
  + Option 1: If new test set up is introduced for AI test
    - Define non-AI requirements under new test set up subject to AI. AI requirement has to be higher than that of non-AI.
  + Option 2: If new test set up is introduced for AI test
    - Define absolute requirement of AI without considering the comparison with non-AI.
  + Option 3: Discuss on a case by case basis
    - there is no need for non AI based requirements if the feature/use case is only enabled with AI
* Recommended WF

To be discussed

### Sub-topic 1-6

*Testing goals*

The following was agreed in the last meeting:

**Issue 1-3: Testing goals**

Agreement:

* The testing goal is to verify whether the minimum performance of AI/ML functionality/feature can be achieved
  + FFS on whether and how many different test configurations/parameters are used for tests.
  + LCM would be tested

**Issue 1-6: RAN4 Testing goals**

* Proposals
  + Option 1: RAN4 to confirm that the new agreement on functionality-based testing goal is substituting Option 1 and Option 2 testing goals considered in Rel-18 and listed in the TR 38.843.
  + Option 2: others
* Recommended WF

Option 1

# 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** |
| [**R4-2414929**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2414929.zip) | MediaTek Inc. | **Observation 1**: Both DL Tx beam prediction and beam prediction are studied in R18 AI/ML SI, but only DL Tx beam prediction is in the scope of R19 AI/ML WI.  **Observation 2**: According to the submitted input from different companies in R18 SI, for DL Tx beam prediction, the model inputs are L1-RSRP of Tx beams in Set B with Tx beam ID information as input or implicitly indicated. The outputs are Probability of each beam in set A to be the Top-1 beam or predicted L1-RSRP of beams in Set A.  **Observation 3**: AoA information is not used at least in all the submitted simulations in R18 SI ([R18 AI/ML BM evaluation in RAN1](https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_114/Inbox/drafts/9.2(FS_NR_AIML_air)/9.2.3.1/BM_eval_results_FINAL/For%20Tdoc%20submission/Table%201.%20Evaluation%20results%20for%20BMCase1_v043_Spreadtrum_Mod.xlsx)). Rx beam information is not used in DL Tx beam prediction.  **Proposal 1**: RAN4 to confirm Rx beam info is not part of the AI/ML model inputs.  **Proposal 2**: The TCI state QCL to an RS that is not in Set B is unknown even if this RS was included in the L1-RSRP report by UE within 1280ms before TCI state activation.  **Proposal 3**: 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.   **Proposal 4**: Discuss how to add the measurement error, i.e., adding error at both training and inference or adding error only on inference. How to add the measurement error shall be aligned with the test goal, i.e., reflect the potential performance of AI/ML BM models in the field or verify UE can train a model that works well under a certain channel and scenario.  **Proposal 5**: Discuss error modelling when defining the prediction accuracy requirements. How to model the measurement error shall be aligned with the test goal, i.e., reflect the potential performance of AI/ML BM models in the field or verify UE can train a model that works well under a certain channel and scenario. |
| [**R4-2415030**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415030.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**   **Observation 1: For option3, when the Top-1 predicted beam isn't the ideal beam, even if the AI/ML prediction accurately predicts the RSRP and achieves an outstanding RSRP accuracy KPI(by 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), it's difficult to assert that the predicted result could be effectively utilized.**  **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.** |
| [**R4-2415071**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415071.zip) | Xiaomi | **Proposal 1: For both BM case-1 and BM case-2, RAN4 to discuss whether there is impact on measurement period due to specific RX beam assumption.**  **Proposal 2: For BM-case 2, RAN4 to discuss impact on measurement requirement, e.g. measurement time, prediction time, measurement restriction and scheduling restriction, etc.**  **Observation 1: 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 3: RAN4 to discuss known TCI state condition for AI based beam indication.**  **Observation 2: For UE-assisted performance monitoring, RAN1 will study how to design evaluation time, evaluation criteria and L1 reporting.**  **Proposal 4: RAN4 to define requirement for UE assistant performance monitoring. The detail needs more input from RAN1.** |
| [**R4-2415170**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415170.zip) | CATT | **Proposal 1: RAN4 postpones the discussions on LCM core requirements until RAN1/2 make sufficient progress.**  **Proposal 2: RAN4 defines unified inference delay requirements for both BM-Case1 and Case2 and discussion is based on BM-Case2.**  **Proposal 3: Inference delay requirements are defined by considering at least the follow parameters:**   * **Periodicity of measurement resources.** * **Periodicity of SMTC/MG.** * **Duration of DRX cycles.** * **F and K (For BM-Case 2).**   **Details are FFS and RAN4 waits for more RAN1 inputs.**  **Proposal 4: For BM cases with NW-side AI/ML model, no AI/ML-specific RRM requirements are needed.**  **Proposal 5: For Type1 network-based performance monitoring, the time duration between the reception of the configuration/signalling for performance monitoring resources and UE reporting can be defined as performance monitoring delay.**  **Proposal 6: For Type2 UE-based performance monitoring, RAN4 to further discuss the necessity of performance monitoring delay requirements.**  **Proposal 7: The TCI state is known if it meets the known conditions defined in current RRM spec. Otherwise, it is unknown.** |
| [**R4-2415190**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415190.zip) | vivo | **Observation 1: There is a non-negligible performance degradation in beam prediction accuracy when considering measurement error modeling in training/validation data.**  **Observation 2: There is no large performance gap between the case that measurement error is only considered in validation data and the case that measurement error is considered both in validation data and training data**  **Proposal 1: RAN4 to define beam prediction accuracy requirements with consideration of the impact on measurement error**  **Observation 3: In addition to the rate of correct beam prediction, when additional 1dB RSRP margin is considered (e.g., For the success rate of the difference between the ideal RSRP of ideal Top-1 beam and the predicted RSRP of predicted Top-K beam are within the X=1dB range), it will bring a noticeable improvement in the rate of prediction accuracy.**  **Observation 4: The probability of predicted RSRP of predicted Top-K beam falling into the range of [the RSRP of Top-1 beam± XdB] has significant improvement with the increase of X value**  **Observation 5: When using predicted RSRP of predicted Top-1 beam and Ideal RSRP of ideal Top-1 beam as the benchmark, the probability that the difference between the RSRP of Top-1 beam and the predicted RSRP of Top-K beam is within the [X]dB range are similar**  **Proposal 2: RAN4 to use Top-K/1 as the metric of beam prediction accuracy for further performance evaluation**  **Observation 6: 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***   **Proposal 3: For metrics/KPIs for Beam Management requirements/tests, RAN4 to align the definition of RSRP accuracy as the difference between the predicted L1-RSRP of predicted beam and the ideal L1-RSRP of the same beam.**  **Proposal 4: For core requirements for beam management, RAN4 to define 1-sample based delay requirements for beam prediction of Spatial-domain in BM case-1. For Temporal beam prediction in BM case-2, the definition of delay requirements may take into account specific methods (e.g., caseA and caseB) and characteristics of time-domain.**  **Observation 7: From the simulated CDF curve of the RSRP difference for modelling errors from LLS and Gaussian distribution, the characteristics of baseband errors (e.g., variance, mean value) derived from LLS and Gaussian distribution are different.**  **Proposal 5: For the LLS-based measurement error modelling, RAN4 to discuss whether the simulation assumption is based on the TR 38.843 or the aligned assumption for L1-RSRP accuracy.**  **Proposal 6: RAN4 to consider the RSRP difference derived from link level simulation under specific SNR condition as the baseband error, and add the error into the ideal dataset derived from system level simulation so as to build the nonideal dataset for training and inference.**  **Observation 8: There is some impact on the prediction accuracy of RSRP when RF error is considered**  **Proposal 7: For the study on the impact of measurement error, the method of RF error modelling needs to be involved in RAN4 discussion, e.g.,**   * **An offset value in the original dataset.** * **Uniform distribution** * **Truncated Gaussian distribution** |
| [**R4-2415249**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415249.zip) | CMCC | ***Proposal 1: the absolute RSRP accuracy is proposed as:***   * ***Absolute RSRP accuracy = predicted L1-RSRP of beam index i – ground truth of beam index i*** * ***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 L1- RSRP on the predicted Tx beam (i.e. transmitted power)***   ***Proposal 2: the relative RSRP accuracy is proposed as:***   * ***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.***   ***Proposal 3: it is proposed that the accuracy requirements apply to all predicted beams.*** |
| [**R4-2415496**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415496.zip) | Apple | **Observation 1: 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 2: 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 3: 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 4: RAN4's goal of aligning test conditions with real-world deployment scenarios is closely related to ensuring generalization performance. Since field deployment conditions can vary significantly across different environments, the test conditions should account for this variability**  **Observation 5: The source of training data for beam management will play a crucial role in AI/ML BM performance and in the generalization performance in real deployment**  **Observation 6:** **For training data based on real measurements, the** **quality of training data depends on RF impairments, and other noise sources. There is tradeoff between training data quality and generalization performance. With training data collection from the field (across varying SNR conditions), both set-B and set-A beams will be affected by impairments**  **Proposal 1**: **RAN4 should investigate the specification of reference AI/ML models for BM, the associated training procedures, and the training datasets to ensure alignment of simulation results. This will help pave the way for defining performance requirements for the BM use case, with careful consideration of constraints on model complexity. To initiate the discussion, we have provided a reference table as a starting point.**  **Proposal 2: 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 3: 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.**  **Proposal 4:** **RAN4 should agree that the training dataset used in testing is strictly for testing purposes. In real-world applications, it will be the responsibility of UE implementations to determine the appropriate training dataset. This could involve using a vendor-specific dataset or creating a mixed dataset sourced from multiple vendors for training their models.**  **Proposal 5: To define the testing data for BM and provide sufficient test environments/conditions for vendors to generate the necessary training data, RAN4 should select a set of standardized test conditions. UEs will be tested across these identified conditions, and it will be up to the UE implementation to either switch between models or employ a supermodel capable of accommodating all these conditions during. Example of test conditions/configurations are:**   * **Details of the channel fading characteristics (CDL, LOS/NLOS, etc)** * **Different paterns of set B/ set A** * **Tx codebooks,Tx antenna architecture layout, antenna spacing** * **SNR conditions** * **Carrier Frequency** * **Doppler conditions, measurement window configurations and window of predicted time instances (BM case 2)**   **Proposal 6: For all potential non-tested conditions that are part of possible deployment scenarios, including additional conditions not explicitly addressed during RAN4 testing, we propose to employ Post-Deployment Procedures. These procedures will augment RAN4 conformance testing to effectively manage and assess performance across all possible deployment conditions**  **Proposal 7: To investigate performance degradation resulting from measurement accuracy, we propose specifying an impairment model as shown 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/Inference 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, correlations between errors for different gNB beams minimum SNR where RSRP is measured, noise figure, etc |   **Proposal 8: 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,**     1. **To maintain performance the number of setp B beams can increase**    2. **The number of K in top-K could increase to compensate for degraded performance. (change model, address scalability of AI/ML model).**    3. **Switch from RSRP prediction to beam prediction with an increased number of beams (K). UE signals the predicted beams to the network (NW) and relies on measurements, rather than predictions, of the NW-transmitted beams identified by the UE.**    4. **Apply AI/ML denoising learning techniques to reject noise and impairments and make the model robust to a range of SNR values** 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.** |
| [**R4-2415624**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415624.zip) | Huawei,HiSilicon | **Proposal 1:**For data collection in NW-sided model, use the existing core requirement for beam related information reporting as the starting point.  **Proposal 2:**If measurement reporting is used for the calculation of performance metric at NW during model monitoring, reuse RAN4 legacy requirement for measurement reporting.  **Proposal 3:**If metric reporting is introduced by RAN1 for model monitoring, RAN4 will study the related requirement. |
| [**R4-2415893**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415893.zip) | Qualcomm Incorporated | **Observation 1: Legacy RAN4 RRM spec allows UE to measure reference signals for a certain duration to meet corresponding accuracy requirements. The required duration depends on SMTC and measurement gap-based sharing factor, and DRX cycle.**  **Observation 2: RAN1 has introduced the concept of association ID that conveys the properties of a DL Tx beam set. UE may require different evaluation period for prediction for different properties of a DL Tx beam set.**  **Observation 3: Option 2’s usefulness lies in the fact that network can use UE’s predicted top beam to change UE’s TCI state.**  **Observation 4: Option 2’s drawback lies in the fact that its pass/fail criteria does not depend on whether the top predicted beams are adjacent to the strongest beams or not. Option 3 does not suffer from this drawback.**  **Observation 5: 3GPP is considering L1-RSRP accuracy as a potential prediction metric for AI-ML based beam management. UE has to know its RX beam to predict L1-RSRP accuracy of a particular gNB TX beam.**  **Observation 6: To compare between gNB beams of setA, UE needs to have some idea of the Rx gain associated with them. UE has to know the Rx beam of the predicted gNB beam even if it only has to predict top-1 (%) or top-K (%) beams.**  **Observation 7: Some companies mentioned that the AI-ML model considered in RAN1 uses (gNB Tx beam, measured RSRP) as inputs and the model does not explicitly indicate UE’s Rx beam. RAN1 assumed that UE knows its corresponding RX beam for the output of the AI-ML model, including predicted gNB Tx beam.**  **Proposal 1: RAN4 should define AI-ML beam management related RRM core requirements in, at least, following three areas:**   1. **Measurement period for prediction** 2. **Accuracy of predicted metrics** 3. **TCI state status of predicted TX beam**   **Proposal 2: Spec allows a minimum duration of measurements on the SSBs/CSI-RS before UE reports a CSI report containing predicted metrics (i.e., channel characteristics).**   * **Note: Measurement period should consider measurement gaps, SMTCs, DRX operation, etc.** * **FFS: Minimum duration may depend on other factors e.g., associated ID, level of accuracy, etc.**   **Proposal 3: Regarding down-selection of prediction metrics, RAN4 takes one of the following two approaches:**   * **Approach 1: Wait for RAN1 to down-select the prediction metrics.** * **Approach 2: Down-select the prediction metrics for RAN4 tests in RAN4.**   + **In approach 2, RAN4 should select option 2 and option 3 of 38.843.**   **Proposal 4: UE knows the RX beam paired with the predicted Tx beam.**   * + **FFS: Whether TCI state should be considered known.**   + **FFS: Whether side conditions for known TCI state should be modified (relaxed) for this case.** |
| [**R4-2416080**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416080.zip) | Ericsson | **Observation 1: It does not make sense to use average L1-RSRP difference of Top-1 predicted beam (referring to TS38.843) for defining RSRP prediction accuracy. ‘RSRP prediction accuracy’ and ‘beam prediction accuracy’ are different concepts, which is exposed in TS38.843 as well.**  **Observation 2: In case of AI/ML model with label (classifier network) output, measurement error may have less impact to the prediction** **than the AI/ML model with RSRP output, if the reporting of the UE only relies on the output of AI/ML model without explicit or implicit assistance of measurement result.**  **Observation 3: UE always knows the Rx beam paired with the predicted Tx beam provided the predicted Tx beam is detectable in training phase and inference phase.**  **Observation 4: In inference phase, UE is able to retain all information relevant to Rx beam (including the association between Tx beam and Rx beam), which is learned from training phase.**  **Observation 5: UE needs not know the Rx beam associated with a Tx beam, if the Tx beam belongs to the below two cases:**   * **Case Set A: the UE has not detected/measured the Tx beam in train phase.** * **Case Set B: the Tx beam is not in Set B.**   **Proposal 1:** **Absolute RSRP accuracy= predicted L1-RSRP of beam index i – ideal measured L1-RSRP of beam index i. i may be any beam index in top-K predicted beams.**  **Proposal 2: The definition of relative RSRP accuracy shall be:**   * **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 value (including predicted and measured beams), if beam index n is not configured for corresponding measurement.** * **Relative RSRP accuracy= predicted L1-RSRP of beam index i – measured L1-RSRP of beam index n, where the beam index n owns the largest value (including predicted and measured beams), if beam index n is configured for corresponding measurement.**   **Proposal 3: Measurement error shall be evaluated with respect to the type of output of AI/ML model, e.g., label (classifier network) or RSRP (CNN or MLP).**  **Proposal 4: Measurement error impact to prediction shall be evaluated for BM-case 1 and BM-case 2 respectively.**  **Proposal 5: Measurement error impact to prediction shall be verified (e.g. by simulation) and the same methodology in Rel-18 RAN1 studies can be referred with the consideration of to guarantee an acceptable prediction performance.**  **Proposal 6: RAN4 to consider the L1-RSRP values for generated channel as the metric in the link-level simulations.**  **Proposal 7: 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 8: RAN4 to consider the following parameters in Table 1 in the link level evaluation**  Table 1 summary of link level evaluation assumptions   |  |  | | --- | --- | | **Parameters** | **Value** | | Frequency range | FR2@30GHz; SCS 120kHz | | Data allocation | 8 RBs  First 2 OFDM symbols for PDCCH, and following 12 OFDM symbols for data channel | | PDCCH decoding | Ideal or non-ideal | | Channel model | LOS channel: CDL-D extension, DS = 100ns  NLOS channel: CDL-A/B/C extension, DS = 100ns | | The measurement inaccuracy | Zero mean Gaussian distribution with 95% of density be within (-6.5, 6.5) dB | | Impairment types affecting set B measurements and set A measurements | 4dB RF impairments with normal distribution, 1dB fading condition  RF impairment model is composed of different elements, including I/Q imbalance, Quantization noise, Phase noise, Filters/Ripple noise, RF PA distortion noise |   **Proposal 9: RAN4 to consider the following parameters in Table 2 in the system level evaluation**  Table 2 Summary of system level evaluation assumptions   |  |  | | --- | --- | | **Parameters** | **Value** | | Frequency range | FR2@30GHz; SCS 120kHz | | Deployment | 200m ISD, 2-tier model with wrap-around (7 sites, 3 sectors/cells per site) | | Scenario | Dense urban (macro-layer only) | | System level channel model | UMa with distance-dependent LoS probability function defined in Table 7.4.2-1 in TR 38.901. | | Number of set -B/ set-A beams, | 8 SSB /32 CSI-RS beams | | Antenna port layout | Antenna setup and port layouts at gNB: (4, 8, 2, 1, 1, 1, 1), (dV, dH) = (0.5, 0.5) λ  Antenna setup and port layouts at UE: (1, 4, 2, 1, 2, 1, 1), 2 panels (left, right) | | Best Rx beam selection algorithm | Baseline 1: Select the best beam within Set A of beams based on the measurement of all RS resources or all possible beams of beam Set A (exhaustive beam sweeping)   Baseline 2: Select the best beam within Set B of beams based on the measurement of all RS resources or all possible beams of beam Set B | | Transmission Power | Maximum Power and Maximum EIRP for base station and UE as given by corresponding scenario in 38.802 (Table A.2.1-1 and Table A.2.1-2) | | Traffic Model | Full buffer |   **Proposal 10: For prediction-based TCI state switch, not to reuse 1280ms as one of known conditions, instead, we shall take training/ inference phase, or training/ inference validity as benchmark in the known conditions.**  **Proposal 11: The known conditions of the predicted TCI state shall take below conditions into account:**   * **The TCI state switch command is received during inference phase of AI/ML mode.** * **The predicted beam in Set A is the RS in target TCI state.** * **The predicted beam in Set A is detectable during training and inference phase, which implies the beam has been trained successfully and available in inference phase.**   **Proposal 12: During training phase, measurement reporting on Set A and Set B also may enable the NW to configure a TCI state associated with a RS out of the measured Set A or Set B. However, it is not relevant to AI/ML model. Hence, the known conditions for Set A or Set B during training phase are not in scope.**  **Proposal 13: During inference phase, it may occur that the TCI state is QCLed with a RS belonging to Set B. The legacy known conditions are applied. Hence, the known conditions for Set B during inference phase are not in scope.**  **Proposal 14: Metrics/KPIs for beam management requirements/tests are separated in below options:**   * **In the case of label (beam ID) output type of AI/ML model**   + **Beam prediction accuracy shall be fulfilled.** * **In the case of label (beam ID) + RSRP output type of AI/ML model**   + **Beam prediction accuracy and RSRP accuracy simultaneously shall be fulfilled.**   **Proposal 15: For BM-Case 1, if a beam in Set A is configured for being measured as well, 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.** |
| [**R4-2416363**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416363.zip) | ZTECorporation,　Sanechips | **Observation 1: Whatever the cost time for UE to report the measurement to gNB, the gNB shall decide when to start to perform the model inference since the model inference will be up to gNB implementation.**  **Observation 2: It is observed that there is no RAN2’s specification impact associated to gNB-side model inference.**  **Proposal 1: From RAN4 perspective, there is no RRM impact for model inference on NW-side and no need to define RRM requirements.**  **Observation 3: After performing the inference, UE shall report the model output (predicted results) to gNB.**  **Observation 4: Two delays shall be considered: (1) model inference delay; (2) UE report delay.**  **Proposal 2: The RRM requirements for model inference of UE-side model shall be considered especially the delay requirements in order to guarantee the accuracy of inference results.**  **Observation 5: RAN2 is still discussing specifications enhancements associated to gNB-side model monitoring.**  **Observation 6: RAN2 deems that the gNB is responsible for monitoring its own performance, which means the gNB-implementation.**  **Proposal 3: RAN4 can wait for the progress of RAN1 and RAN2, then we can decide whether the RRM requirements shall be defined or not.**  **Observation 7: For Type 2 performance monitoring, RAN1 deems that the indication/request/report from UE to gNB for performance monitoring is not needed.**  **Observation 8: For Type 1 performance monitoring, there are two options: (1) NW-side performance monitoring, UE sends the report to gNB; (2) UE-assisted performance monitoring, UE calculates the performance metrics and reports it to gNB.**  **Observation 9: For Type 1 performance monitoring, option 1 and option 2 are similar to the inference on UE-side model.**  **Proposal 4: The performance monitoring will be up to UE implementation and there is no RRM impact on Type 2 performance monitoring.**  **Proposal 5: RAN4 shall consider to define the RRM requirements for Type 1 performance monitoring like inference on UE-side if it has.** |
| [**R4-2416439**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416439.zip) | Nokia | In this paper we share our views on potential RAN4 impacts from issues related to RRM core requirements for AI/ML enabled BM-Case1 and BM-Case2. Specifically, we cover following aspects:   * Measurement error impact and the approximated ground truth * Indicated/activated TCI states requirements * Impacts on requirements related to candidate beam detection * KPIs for beam prediction * LCM related requirements   In the paper, the following Observations and Proposals were made:  **Observation 1:** The AI/ML method is very sensitive to measurement error. The RSRP difference between measurement with error (±6dB) and measurement without error is 5 dB higher than non-ML, suggesting that the performance of the AI/ML method degrades significantly with legacy RSRP measurement error requirement.  **Observation 2:** To align the AI/ML method's performance with the non-AI/ML approach, it is reasonable to consider reducing the measurement error margins. Specifically, reducing the measurement error to ±4 dB or ±3 dB result in ΔeRSRP values of 3.68 dB and 2.07 dB, respectively, which are more comparable to the non-AI/ML performance.  **Proposal 1: Given the significant degradation in prediction performance observed with the current ±6 dB measurement error margin, RAN4 shall discuss:**  **i) whether such degradation (around 5 dB in simulations) with respect to legacy performance is acceptable or not.**  **ii) whether impacts on other performance metrics should be studied or not**  **Proposal 2: To ensure that the AI/ML method performs comparably to the non-AI/ML method, RAN4 shall discuss:**  **i) whether a tightening in the legacy measurements’ requirement is feasible/acceptable.**  **ii) If there are other ways for reducing the gap between AI/ML and non-AI/ML prediction performance in presence of measurements error**  **Proposal 3: RAN4 should discuss if filtered L1-RSRP measurements can be alternative to improve prediction accuracy and can be used to avoid any tightening in legacy measurement accuracy requirements.**  **Proposal 4: It is reasonable to assume that UE may know the Rx beam to be used for receiving the predicted Tx beam, therefore Option 2 should be selected.**  **Proposal 5: RAN4 should consider relaxing known conditions for TCI states (Clause 8.10.2 of TS 38.133) for BM-Case1 and BM-Case2.**  **Proposal 6: RAN4 should consider the impact in RAN4 requirements for the Evaluation period**  **due to the presence of predicted beams.**  **Proposal 7: RAN4 should define RSRP accuracy metric per beam as, “The accuracy for the predicted L1-RSRP of Top-K predicted beams should be within a specified margin with respect to the ground truth for corresponding beams”.**  **Observation 3:** Top-1(%) metric of Option2 is the strictest requirement and RAN4 should not consider it as a performance metric.  **Observation 4:** Top-K/1(%) metric of Option2 only verifies if actual strongest beam is among Top-K predicted beams and it doesn’t verify the accuracy of prediction for other beams.  **Proposal 8: Top-K/1(%) metric of Option 2 should be considered in the combination of Top-1/K(%).**  **Observation 5:** Top-1/K(%) metric of Option2 only verifies if Top-1 predicted beam is among Top-K actual strongest beams and it doesn’t verify the accuracy of other predictions.  **Observation 6:** Option 3 only compares the measured L1-RSRP of the strongest predicted beam with the measured L1-RSRP of the actual strongest beam. It does not verify the quality of prediction for other predicted beams.  **Proposal 9: RAN4 should consider Option 3 but in combination of other Options (such as Option 1 and Option 2).**  **Observation 7:** The behaviour of beam prediction functionalities can be tested without explicit definition of the ground truth by comparing the predictions to each other.  **Proposal 10: 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.**  **Proposal 11: Option 5 should be considered in the combination of other approximated ground truth-based options (such as Option 1).**  **Proposal 12: The delay components of LCM operations should be considered for different CSI-report configurations (i) periodic CSI report (ii) semi-persistent CSI report and (iii) aperiodic CSI report for both BM-Case1 and BM-Case2.**  **Proposal 13: RAN4 should consider Table below as a baseline to start discussions for LCM operations related RRM core requirements.**  **Table: Delay components related LCM operations for different CSI-report configurations**   |  |  |  |  | | --- | --- | --- | --- | | **CSI-Report Configuration** | **Activation** | **Switching** | **Fall-back**  **(From AI/ML BM to legacy)** | | **Periodic CSI-reporting** | * Delay from RRC Reconfiguration * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay from RRC Reconfiguration * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay from RRC Reconfiguration * Delay from UE processing * Delay due to CSI-RS measurement period | | **Semi-persistent CSI-reporting** | * Delay for activating semi-persistent CSI resource set * Delay for selecting of semi-persistent CSI resource set * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay for selecting of semi-persistent CSI resource set * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay for activating semi-persistent CSI resource set * Delay for selecting of semi-persistent CSI resource set * Delay from UE processing * Delay due to CSI-RS measurement period | | **Aperiodic CSI reporting** | * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay from UE processing * Delay related to associated ID * Delay due to CSI-RS measurement period | * Delay from UE processing * Delay due to CSI-RS measurement period | |
| [**R4-2416492**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416492.zip) | Samsung | *General handling for AI-BM core requirement*  **Proposal 1:** For AI-BM core requirement, RAN4 shall firstly agree on the expected RAN4 requirement impacts from the perspectives of (1) NW and UE-sided model; (2) different sub-features introduced in RAN1/2, such as: Data collection, Inference, Perf. monitoring, LCM and additional assistance information (for UE-sided model only).  **Proposal 2:** RAN4 use the below table as the template for collecting views from companies, on RAN4 core requirement impact for AI-BM use case.   |  |  |  |  | | --- | --- | --- | --- | |  | New Sub-Feature introduced in RAN1/2 | Whether RAN4 requirement impact is expected? | Clause in TS 38.133 impacted, if any | | NW-sided model | Data collection (for inference and/or training) | *e.g., Yes or No* | *e.g., new clause or clause x.x or no impact* | | Inference |  |  | | Perf. monitoring |  |  | | LCM |  |  | | UE- sided model | Data collection for training |  |  | | Inference (including additional assistance information) |  |  | | Perf. monitoring |  |  | | LCM |  |  |   *RAN4 Core Requirements for Supporting NW-sided Model*  **Observation 1:** For NW-sided model, for inference report, enhancement on beam reporting in L1 signaling is agreed to be introduce to accommodate Top-M (where M can be larger than 4) beams with corresponding beam information.  **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:   * L1 measurement/reporting for inference: e.g., Top-M (where M can be larger than 4) beams with corresponding beam information;   **Proposal 4:** No RAN4 impact is expected for NW-side data collection.  **Observation 2:** RAN1 has studied the impact of measurement error on prediction accuracy (1) by modeling BB and RF factors as combined or differentiated errors; (2) by studying the different level of relative measurement error; and (3) by studying the impact for DL TX beam predication and TX-RX beam pair prediction.  **Observation 3:** Lower prediction accuracy is expected for both DL TX beam predication and TX-RX beam pair prediction, but no clear alignment is achieved among companies’ results from RAN1 evaluation.  **Proposal 5:** For measurement accuracy requirement for AI-BM NW-sided model:   * Unless the feasibility of tightened measurement accuracy is demonstrated, RAN4 shall reuse the accuracy requirement from Rel-15 L1-RSRP for both absolute and relative RSRP reporting.   **Proposal 6:** No RAN4 impact is expected for NW-side AI/ML model inference for BM-Case1 & 2, except data collection for model inference.  **Proposal 7:** FFS RAN4 impact from NW-side AI/ML model performance monitoring for BM-Case1 & 2 (if RAN1/2 introduce any mechanism for UE side additional conditions)  **Proposal 8:** No RAN4 impact is expected for NW-side AI/ML model LCM for BM-Case1 & 2.  **Proposal 9:** Based on corresponding proposals, the expected RAN4 requirement impacts from AI-BM NW-sided model are summarized as:   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **New Sub-Feature introduced in RAN1/2** | **Whether RAN4 requirement impact is expected?** | **Clause in TS 38.133 impacted, if any** | **Related Proposal(s)** | | **NW-sided model** | Data collection (for inference and/or training) | Yes: for L1-RSRP reporting  No: for NW-side data collection | • Clause 9.5 L1-RSRP measurements for reporting (or similar new requirement for L1-RSRP measurement for reporting for AI-BM) | Proposal 3,4,5 | | Inference | No |  | Proposal 6 | | Perf. monitoring | FFS RAN4 impact |  | Proposal 7 | | LCM | No |  | Proposal 8 |   *RAN4 Core Requirements for Supporting UE-sided Model*  **Proposal 10:** No RAN4 impact is expected from data collection for UE-sided model training.  **Proposal 11:** RAN4 requirement of UE-side AI/ML model inference shall be specified.  **Proposal 12:** RAN4 shall specify RRM requirement for UE-side AI/ML model inference: to guarantee (1) accuracy of inference results accordingly to the agreed KPI; (2) the delay of reporting inference results (containing the inference latency).  **Observation 2:** For UE-sided model, DL Tx beam prediction are considered with the measurements from (1) the best Rx beam or (2) Tx-Rx beam pair, as the input to AI/ML network.  **Proposal 13:** 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 by applying the same associated ID as side condition.  **Proposal 14:** 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): No RAN4 requirement is required on data collection for monitoring, because it is similar to data collection for other purposes. 2. Type 1, Option 2 (UE-assisted performance monitoring): FFS until RAN1 agreement on different alternatives. At least for some of alternatives, RAN4 prediction accuracy requirement can be used to guarantee this kind of performance monitoring. 3. Type 2 (UE-side performance monitoring): No RAN4 requirement is required because no UE feedback will be performed and the performance monitoring can be guarantee by model generalization test implicitly.   **Proposal 15:** 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).   **Proposal 16:** Based on corresponding proposals, the expected RAN4 requirement impacts from AI-BM UE-sided model are summarized as:   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **New Sub-Feature introduced in RAN1/2** | **Whether RAN4 requirement impact is expected?** | **Clause in TS 38.133 impacted, if any** | **Related Proposal(s)** | | **UE-sided model** | Data collection for training | No |  | Proposal 10 | | Inference (including additional assistance information) | Yes | New clause for UE-side AI/ML model inference (to guarantee (1) prediction accuracy; (2) the delay of reporting inference results) | Proposal 11-13 | | Perf. monitoring | No: For Type 1 (Option 1) and Type 2  FFS: Type 1 (Option 2) |  | Proposal 14 | | LCM | Yes | FFS detailed RAN4 requirement | Proposal 15 | |
| [**R4-2414930**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2414930.zip) | MediaTek Inc. | **Observation 1**: Both DL Tx beam prediction and beam prediction are studied in R18 AI/ML SI, but only DL Tx beam prediction is in the scope of R19 AI/ML WI.  **Observation 2**: According to the submitted input from different companies in R18 SI, for DL Tx beam prediction, the model inputs are L1-RSRP of Tx beams in Set B with Tx beam ID information as input or implicitly indicated. The outputs are Probability of each beam in set A to be the Top-1 beam or predicted L1-RSRP of beams in Set A.  **Observation 3**: AoA information is not used at least in all the submitted simulations in R18 SI ([R18 AI/ML BM evaluation in RAN1](https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_114/Inbox/drafts/9.2(FS_NR_AIML_air)/9.2.3.1/BM_eval_results_FINAL/For%20Tdoc%20submission/Table%201.%20Evaluation%20results%20for%20BMCase1_v043_Spreadtrum_Mod.xlsx)). Rx beam information is not used in DL Tx beam prediction.  **Proposal 1**: If to use multiple AoAs test system for AI/ML BM test, test system should be able to emulate as many Tx beams as in set A, that is at least 64.  **Observation 4**: It is difficult to ensure reference RSRP is measured under similar channel conditions as in prediction phase.  **Proposal 2**: 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 3**: If to use multiple AoAs test system for AI/ML BM test, UE rotation should be supported. Otherwise, Rx beamforming gain for each AoA will be basically fixed during the test.  **Proposal 4**: RAN4 should further discuss how to handle potential performance loss if simplified CDL channel is used in AI/ML BM test.  **Observation 5**: Designing tests based on CDL channel may not guarantee the real field performance, because AI/ML models are quite scenario/configuration sensitive.  **Proposal 5**: It is not necessary to emulate AoAs and AoDs of each Tx beam as the information is not used in all the submitted simulations in R18 SI.  **Proposal 6**: In AI/ML BM test, emulate the spatial-selective and time-varying in propagation conditions through adjusting Tx power at TE on each beam at different time.  **Observation 6**: Single AoA test system can be applicable to all kinds of synthetic channels as well as using real field data. It can be easily extended to different scenarios and settings.  **Observation 7**: With single AoA, TE can know the expected RSRP of Tx beams at UE baseband.  **Proposal 7**: Use existing IFF or enhance IFF test systems with single AoA for R19 AI/ML BM test. |
| [**R4-2414960**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2414960.zip) | CAICT | **Proposal 1: Consider to define a more stringent requirements for ideal measurement used to generate ground truth and discuss how to verify the ideal measurement performance.**  **Proposal 2: If following the definition of relative RSRP accuracy for prediction in option 1, legacy relative RSRP accuracy requirements could be considered as starting point.**  **Proposal 3: Suggest to add a new RSRP accuracy metric, specifically, the ideal measured L1-RSRP of beam index i – ideal measured L1-RSRP of beam index k, where the beam index i owns the largest predicted value and beam index k owns the largest measured value.** |
| [**R4-2415031**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415031.zip) | OPPO | **Proposal 1: 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 2:**   * **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 3: 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.**  **Proposal 4: The simplified CDL should be used as the channel assumption.** |
| [**R4-2415070**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415070.zip) | Xiaomi | **Proposal 1: For simplified CDL model, for each cluster, only one ray is considered.**  **Proposal 2: For simplified CDL model, cluster number can be reduced from spatial angle offset and power aspect:**   * **Combine clusters if AOA offset between these clusters is smaller than a threshold** * **Remove cluster with low power.**   **Proposal 3: Measurement error may have impact on prediction accuracy requirement and performance metric selection.**  **Observation 1: In RAN1, the measurement error is modeled as Gaussian distribution for RF and BB, where the error range is 2dB, 4dB and 6dB.**  **Observation 2: In RAN1, the RSRP prediction degradation is calculated based on different assumptions and there is no aligned comparison. It’s difficult for RAN4 to directly use the results.**  **Observation 3: When label for training and test is not aligned, Performance degradation is expected.**  **Observation 4: When label for training and test is aligned, whether there will be performance degradation depends on whether measurement error for input data is aligned or not. Measurement error difference may be impacted by SNR and error statistic distribution assumption.**  **Proposal 4: RAN4 need to first align the assumptions of input data and labels for training and test first, then further analyze performance degradation due to impact of measurement error if needed.**  **Observation 5: According to the latest RAN1 agreement, the inference report content can include beam information and predicted RSRP.**  **Observation 6: The beam information is similar to the legacy beam index reporting and is presented in the form of an RS ID configured in a resource set.**  **Proposal 5: RAN4 to start to discuss the refinement of RSRP accuracy definition.**  **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.** **Relative RSRP accuracy is the RSRP difference between the predicted L1-RSRP of different beams.**   **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 7: In legacy non-AI BM, both beam index and RSRP can be included in L1 report, however, RAN4 only defines requirement for RSRP accuracy and RAN4 doesn’t define requirement for beam accuracy.**  **Observation 8: For AI, Whether beam accuracy metric can be skipped depends on how accurate absolute RSRP prediction can achieve.**  **Observation 9: From TR, predicted best TX beam index and/or L1-RSRP can be based on full RX beam sweeping or specific RX beam.**  **Observation 10: 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 11: 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 12: 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 8: RAN4 to discuss how to keep the RX beam Consistency between prediction and measurement for set A to correctly evaluate prediction performance.**  **Proposal 9: Suggest to use RSRP delta CDF-based simulation to determine the appropriate SNR level for deriving ground truth.**  **Proposal 10: For BM case-1, channel doppler can set to 0 or a small value to guarantee that there is neglectable L1-RSRP variation.**  **Proposal 11: 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 12: 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 13: UE will report best predicted L1-RSRP/beam index at T1 and measured RSRP/beam index of best beam at T2 to TE, it’s easy for UE to pass the test by cheating.**  **Proposal 13: RAN4 to discuss how to solve the UE cheating issue if UE report both predicted result and ground truth.** |
| [**R4-2415171**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415171.zip) | CATT | **Proposal 1: Relative RSRP accuracy is modified as follows:**   * + - **Relative RSRP accuracy = predicted L1-RSRP of beam index i – ~~predicted~~ measured L1-RSRP of predicted beam index n, where the beam index n owns the largest ~~predicted~~ measured value obtained by measuring the predicted beams.**   **Proposal 2: The definition of absolute RSRP accuracy can be used as test metric directly. For relative RSRP accuracy metric, it can be defined as:**   * + - **The relative RSRP accuracy shall be less than X dB, where X are relative accuracy requirements and will be defined in performance part according to the agreed definition.**   **Proposal 3: For beam prediction accuracy, RAN4 waits for RAN1 progress in monitoring metric which could be reused for beam prediction tests directly.**  **Proposal 4: Measurement errors can be modelled with a certain statistical distribution, e.g., uniform or Gaussian distribution, and added to ideal measurement results. Then the impacts can be evaluated by simulations.** |
| [**R4-2415191**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415191.zip) | vivo | ***Observation 1: The simplified CDL channel with small number clusters will degrade the test dependability.***  ***Observation 2: 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 3: Different test system with reduced cost and complexity can be introduced by using the Tx beam grouping method.***  ***Proposal 1: RAN4 to consider methods to group Tx beams together to reduce the number of probes in the test system.***  ***Proposal 2: RAN4 to consider the following procedures to simplify the test system:***   * ***Step 1. Generate the BS beamforming power pattern of every Tx beam*** * ***Step 2. Add the BS beamforming power pattern into the power of different clusters in specific ZoD and AoD, and derive the new power of every cluster*** * ***Step 3. Use the new power of clusters to generate the power angular spectrum (PAS) of ZoA and AoA*** * ***Step 4. Some Tx beams with the corresponding ZoA/AoA are to be grouped together based on a certain rule (e.g., the ZoA/AoA of the strongest power derived from the PAS in step 3 are the same angle or in a pre-defined range)*** * ***Step 5. Using one or small number of probes to emulate one beam group.*** * ***Step 6. Delete the unnecessary probes and merge some adjacent probes.***   ***Proposal 3: RAN4 to consider the following two approaches for design the test system***    ***approach a***    ***approach b***  ***Proposal 4: RAN4 to come back the test setup needs issue (number of beams, AoA, AoD, UE rotation, etc) after noticeable progress of channel model and test framework issues are made.*** |
| [**R4-2415248**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415248.zip) | CMCC | ***Proposal 1: 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 2: 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.*** |
| [**R4-2415312**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415312.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: Verifying the AI/ML beam management performance (e.g., prediction of top K beams) in a TDL channel without spatial characteristics makes the test too simple and inconsistent with the actual field environment.**  **Observation 3: Some companies suggested to emulate multiple beams with single AoA by using different powers. The power levels will have to be defined in test configuration to enable this method. That will allow UE to predict the strongest beam through simple interpolation.**  **Observation 4: CDL with spatial properties is a realistic and proper channel model for AI/ML beam management testing.**  **Observation 5: The key requirements for AI/ML BM test setup are listed following:**   * **Requirement#1: Channel model in the test**   + **CDL or simplified CDL is needed for multiple AoAs to properly verify beam management performance.** * **Requirement#2: Avoid the cheating**   + **Progress A: Data training for Set A’s is generated with channel parameters #1. Set B in UE measurement should have enough randomness compared to Set A’.**   + **Progress B: Set A is generated to make sure to use the same parameters (initial phase, angular spread, delay spread) as the Set B in the Progress A measurement step. NW antenna/beam configurations for Set A and Set B should be deterministic.** * **Requirement#3: Support the beam number of Set A**   + **TE needs to support Set A for ground truth verification, i.e., upto 128.**   **Observation 6: There will be plenty of additional probes if BS Tx beams point to other clusters even though some of the probes could be reused.**  **Observation 7: The UE’s antenna gain gap between peak and 50%-tile spherical coverage is over 10dB for PC3. The very weak clusters would not influence the UE beam manamgent even if they reach the UE within the UE’s spherical coverage.**  **Observation 8: The AoDs/ZoDs of very weak clusters are unlikely to influence the selection of the strongest beam in any reasonably designed codebook of setA and setB.**  **Observation 9: The weak clusters that have a limited impact on the beam management could be removed to reduce the number of clusters in the CDL channel model.**  **Observation 10: The clusters that have the same/similar AoAs could be merged. And the intra-cluster angle spread, e.g., CASD, CASA, CZSD, CZSA can be reduced to simplify the test setup.**  **Observation 11: TE vendors need to check UE positioning time in order to mimic CDL channel model by sequentially rotating the positioner. Order of ~1s dwell time will lead to an unrealistic test environment for AI/ML beam management testing.**  **Observation 12: For some prediction metrics (e.g., option 2 and 3), UE can pass the test without actual measurement if TE does not randomize the transmission of its beams during measurement step of progress of Figure 3.**  **Proposal 1: Take FR2 MIMO OTA as the baseline test setup for AI/ML BM testing and further discuss how to enhance the FR2 MIMO OTA test setup to support the multiple AoA/AoD (e.g., beam number of Set A) under simplified CDL channel models with enough randomness.**  **Proposal 2: RAN4 considers reducing the cluster number with weak power level, merging the cluster with the same/similar AoA, and decreasing intra-cluster angle spread in CDL channel model to simplify the test setup based on the evaluation of the impact on beam management performance considering the UE beam width.**  **Proposal 3: During measurement step of progressB, TE randomizes the transmission of setA beams, (e.g., during time t1, TE does not transmit with beam shape A1, and TE transmits using the shape of a beam whose ID is randomly selected from set A).** |
| [**R4-2415497**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415497.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: For RAN 4 testing we should have the following goals: (1) Provide confidence that if a DUT passes the test it will also perform well in the field (2) Ensure reliability of using synthetic channels for test data in evaluating models trained on real data.**  **Observation 6: (BM testing set up):**  **Advantages (Option 5): The channel model captures simultaneous reception across multiple AoAs/AoDs, providing a more accurate representation of CDL channels and real-world field data.**  **Disadvantages (Option 5): It may lead to a more complex testing setup. There could be challenges in selecting the appropriate simplified CDL channel.**  **Advantages (Option 6): The testing setup is simplified, and existing OTA set up can be reused. For beam prediction only (not RSRP prediction) it may be possible to avoid transmission of set A beams since TE could know the best beam before hand.**  **Disadvantages (Option 6): It may not accurately represent real field conditions. The reliability of using a simplified channel model for testing data may be questionable when evaluating models trained on real-world data.**  **Observation 7:** **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 8:** **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 9: 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 10: Achieving consistency between training and inference by model monitoring could result in delays and increased complexity in model management for BM use case**  **Observation 11: 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 12: 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.** ***Beam Prediction Testability Discussion (Testing set-up)*** **Proposal 1: Multipath based Testing Setup for BM: Evaluate the testability requirements for simulating time-varying input power to a sparse probe layout based on a simplified channel CDL model. The key aspects to be considered are:**   1. **How to determine the minimum number of clusters to be emulated without introducing bias in the results. Investigate a quantitative “goodness” criterion to selecting the number of probes/clusters.** 2. **Consider incorporating UE rotation for the test** 3. **How many different CDL channels we need to support (considering generalization purposes as well)**   **Proposal 2: To establish a criterion for assessing the feasibility of a simplified sparse probe layout (Option 5 simplified CDL) for beam management (BM) testing, the following procedure can be considered:**   1. **Train a model using the reference CDL channel and test it with the simplified sparse layout-based CDL channel.** 2. **Train a model using the simplified CDL channel and test it with the reference CDL channel model.** 3. **Train and test a model using the reference CDL channel.**   **A similarity metric should be defined to compare the RSRP accuracies across these test cases. This metric would indicate the effectiveness of testing with the simplified CDL model. The similarity metric can be derived through simulations. Below figure: CDL\_ref is the reference channel to be approximated by the simplified CDL\_TE channel during test.**  **A diagram of a training  Description automatically generated**  **Proposal 3: For BM-Case 1 spatial prediction and for verifying RSRP accuracy we propose to use the testing setup and channel emulator functionality as described in Figure below. (multiple static configurations) (Option 5)**  **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 4: For testing the beam prediction/RSRP accuracy for BM-Case 2 temporal prediction we propose the framework described in this section and visualized in Figure below, through a non static configuration of the channel emulator. (dynamic channel)**  A diagram of a computer program  Description automatically generated with medium confidence  **Proposal 5: For BM spatial prediction RAN4 to consider the following options option 5 and option 6**  **Option 5 Multipath based Testing Setup with a simplified CDL channel and probe layout. Criterion of evaluating the feasibility of a particular simplified CDL model has been proposed.**  **Option 6 Single path/cluster (LOS condition) based Testing Setup**  **Option 6 for BM testing set-up is shown below:**  A diagram of a circle with lines and dots  Description automatically generated with medium confidence  ***Generalization issues for BM***  **Proposal 6: 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 7: 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 8: 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 9:** **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 10: 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 11: RAN4 to consider option 1 for achieving consistency between training and inference for only static conditions.**  **Proposal 12:  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 13: 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 14: 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 15: 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 16: 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.**  **A diagram of a model  Description automatically generated**  **Proposal 17: 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 18: 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**  A diagram of a machine learning  Description automatically generated  **Proposal 19: 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.** |
| [**R4-2415625**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415625.zip) | Huawei,HiSilicon | ***Observation 1:*** AoA and AoD are not used as model input in AI BM Case 1 and Case 2.  ***Proposal 1:*** Use TDL as the channel model for AI BM testing.  ***Proposal 2:*** To consider the measurement error impact, the relationship between SNR and RSRP measurement error can be used to generate model input during model training.  ***Proposal 3:*** 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 avoid that a UE can pass the test but performs 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 4:*** 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. |
| [**R4-2416081**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416081.zip) | Ericsson | **Observation 1: The achievable high SNR without distorting to measurement in test environment shall be studied. Test environment vendors may provide more information on it.**  **Observation 2: Error (offset) between predicted RSRP and ideal measurement RSRP may depend on whether the two errors are positive correlated, negative correlated or noncorrelated.**  **Observation 3: The prediction accuracy in filed deployment is decided or dominated by the union of AI/ML model and other parts in UE, not only AI/ML model.**  **Observation 4: A simplified AWGN channel together with power changed beams may be too simple to reflect the expected effect of conformance test, instead, it more like a function test for AI/ML model, which cannot reflect field deployment conditions.**  **Observation 5: A CDL channel reflects more accurate field deployment conditions.**  **Proposal 1: Test procedure shall at least comprise the below three steps:**   * **Acquire the predicted RSRP under the required corresponding SNR condition** * **Acquire the ideal measurement of RSRP under high SNR condition** * **Compare the two RSRPs acquired in the above steps**   **Arranging the details on the sequence and test configuration may need further investigation.**  **Proposal 2: To derive SNR condition, simulation study with agreed simulation configurations is expected.**  **Proposal 3: UE under test is able to perform RSRP measurement and prediction with the same test configurations but different SNR conditions.**  **Proposal 4: How to avoid cheating in the reports on RSRP measurement and prediction is a valid issue, which shall be studied. A simple way for the test environment is adding a procedure manually switch the beam indexes between the prediction duration and the measurement duration which is unknown to UE under test.**  **Proposal 5: RAN4 shall identify the test purpose before selecting the channel model.**   * **If the test is to imitate field deployment conditions, a CDL channel is preferred.** * **If the test only examines the function of AI/ML model, a simplified AWGN channel together with power varying signal through OTA may work.**   **We prefer CDL channel, as it imitates the field deployment conditions, which is aligned with our target.**  **Proposal 6: RAN4 to agree that the maximal number of Set A/Set B is the maximal number of Set A/Set B that a UE supports mandatorily, and the exact number is left for RAN1 to decide.**  **Proposal 7: For BM-case1, at least two sets of number of Set A/Set B are provided for test. Below are two examples:**   * **For verifying prediction model computational power/complexity, one example is that number of set A = X, number of set B = X/4.** * **For verifying prediction scaling, one example is that number of set A = X, number of set B = X/16**   **Different accuracy requirements may be applied for different sets of number of set A and set B.**  **Where X may be the maximum number of Set A that a UE supports mandatorily.**  **Proposal 8: For BM-case2, the number of set A and set B may be the maximum number of set A and set B that a UE supports mandatorily.**  **Proposal 9: RAN4 to define the number of future time instances to predict, i.e., prediction window, for BM-case 2.**  **Proposal 10: Take 3D Multi-Probe Anechoic Chamber (MPAC) for FR2 MIMO OTA testing as the start point of reference methodology 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, generating 2 channel instants including channel #1 for training and generating channel #2 with same parameters for inference.** * **For BM-case2, plural channel instants with a series of changes of the channel parameters are necessary to reflect temporal correlation between channels. At least, it includes:**   + **Generating a set of channel instants, from channel #1 to channel #m, for training, and**   + **Generating a set of channel instants, from channel #m+1 to channel #n for inference**   **Wherein,**   * + **The channels instants comprise sequent and grade updates of a set of parameters.**   + **The numbers of m and n depend on the measurement window and prediction window of BM-case2.** * **The channel parameters may comprise one or more than one of**    + **Departure angles (AOD, ZOD)**   + **Arrival angles (AOA, ZOA)**   + **UE position and/or rotation**   + **Channel coefficient**   + **Others aren’t precluded**   **Proposal 12: 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 and match the practical field deployment conditions.** |
| [**R4-2416362**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416362.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: CDL can reflect the characteristic of spatial domain. In order to reflect the real-word environment, CDL shall be considered instead of TDL models since TDL models cannot reflect critical beamforming effects at all.**  **Observation 9: However, based on TR38.901 and TR 38.827, there are a lot of clusters for CDL, 24 clusters which are to many for AI/ML beam prediction testing. It is not practical to use all of the clusters without change since it will have the heavy complexity and cost of the test system.**  **Proposal 6: RAN4 shall support the simplified CDL models.**  **Proposal 7: For BM case 1, two CDL channels are enough. For BM case 2, the tracking change shall be considered if RAN4 would like to reduce the CDL clusters.** |
| [**R4-2416391**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416391.zip) | ROHDE & SCHWARZ | **Observation 1:** Because of the OTA environment and multi-dimensional aspects of AI/ML BM, the test system can become extremely complex, hardly maintainable and commercially unfeasible.  **Observation 2:** AI/ML BM testing needs to be covered in a test system with broader coverage in terms of other features and lifetime, with reasonable complexity, which makes it affordable also to vendors not targeting AI/ML BM testing.  **Observation 3:** AI/ML BM testability discussion started based on the assumption that CDL channel model are relevant for AI/ML BM core requirement and that the MIMO-OTA system supports such testing.  **Observation 4:** Detailed testability discussions showed that upscaling the MIMO OTA system for the AI/ML BM scenarios is not feasible.  **Observation 5:** Detailed requirement discussions showed that there is not yet a common understanding among vendors about the necessity of CDL channel and about the overall (including all aspects) scope of the AI/ML BM testing.  **Proposal 1:** RAN4 to first converge to a common understanding and expectation about the AI/ML BM core requirements. Only then, a testability discussion can be conducted efficiently and goal-oriented.  **Observation 6:** There are many aspects which need to be discussed and concluded whether they are relevant and need to be included in the AI/ML BM requirements or not.  **Question 1:** Do the AI/ML BM requirements focus only on the performance of the AI/ML model, which deals only with L1-RSRP values (input: Set B beams L1-RSRP, output: predicted Set A beams L1-RSRP), or go beyond that, including the performance of other aspects?  **Observation 7:** If AI/ML BM requirements focus only on AI/ML model performance, the requirement for the test system are the less complex ones.  **Question 2:** Do the AI/ML BM requirements include the L1-RSRP UE measurement performance?  **Observation 8:** If L1-RSRP UE measurement performance is included, large accuracy tolerances might loosen the AI/ML performance, reference antenna gain might be required and CATR technology in unavoidable.  **Question 3:** Do the AI/ML BM requirements include UE Rx beam sweeping performance?  **Observation 9:** If UE Rx beam sweeping performance is not included, single AoA testing might be sufficient.  **Question 4:** How many Set B beams are required? Are the prediction results validated against emulated Set A beams? If yes, how many Set A beams are emulated?  **Observation 10:** From a testability perspective the number of Tx beams become very critical, if combined with directivity aspects and spatial (multi AoA) channel models.  **Question 5:** Do the AI/ML BM requirements include the performance under multipath fading channel model? If yes, of which type?  **Observation 11:** It is not clear how to test real performance AI/ML models with testing-oriented (synthetic) channel models.  **Question 6:** Do the AI/ML BM requirements include UE directivity performance and performance under spatial (multi AoA) multipath fading channel models?  **Observation 12:** Support of spatial (multi AoA) channels goes on the cost of reducing extremely the number of Set B/A beams, which in fact underperform the intended AI/ML model, which is the main test objective.  **Proposal 2:** The AI/ML BM requirements should focus on the essential KPIs impacting the performance of AI/ML model.  **Observation 13:** Including too many aspects in the AI/ML BM requirement, which are tested also elsewhere separately, diverge the testing scope to an end-to-end test, risking to leave the key component, AI/ML model, not thoroughly tested. |
| [**R4-2416440**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416440.zip) | Nokia | **Observation 1:** A functionality, if tested in different generalization scenarios, would be considered not overfitted to a certain test environment.  **Proposal 1: Testing of generalization aspects should also be considered to avoid overfitting of functionalities to the test environment.**  **Proposal 2: 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.**  Table: Parameters for Generalization Scenarios   |  |  | | --- | --- | | **Parameters** | **Description** | | UE Speed | Slow / Medium / Fast | | SINR (Deployment Scenario) | Good / Bad Radio conditions | | Carrier Frequency | 2GHz / 3.5GHz | | Propagation Model | TDL / CDL |   **Proposal 3: RAN4 needs to design a new metric, indicative of generalization capabilities of AI/ML BM functionality, to verify the generalization performance of the functionality in different scenarios.**  **Proposal 4: RAN4 to consider the advantages and drawbacks of each of the CDL and TDL channel models before deciding the channel model for the conformance testing of AI/ML-enabled BM use case.**  **Observation 2:** CDL channel models present a more realistic way for testing AI/ML BM use cases due to the more realistic spatial behaviour.  **Observation 3:** If the testability objective of AI/ML-enabled BM use case is to verify the AI/ML BM functionality taking only L1-RSRP of Set B beams as input and not to test the multipath reception of the UE, the use of TDL channel model during the conformance testing is sufficient.  **Observation 4:** The clusters of ‘CDL A’ channel model spans the widest range in the zenith domain whereas the range of all CDL channel models in azimuth domain is almost similar.  **Observation 5:** The clusters of ‘CDL D’ channel model spans the widest range in the azimuth domain while the spans of the clusters of ‘CDL D’ and ‘CDL E’ channel models in the zenith domain are almost same and is narrowest relative to the other CDL models.  **Observation 6:** As the FR2 channel is spatially sparse in nature, simplified CDL model with reduced number of clusters will be a good approach for the conformance testing of AI/ML-enabled BM use-case.  **Proposal 5: RAN4 to decide the number of clusters and their angular location in the simplified CDL model considering the following aspects:**   1. **Practical number of test probes to be made available in the OTA test chamber for AI/ML-enabled BM conformance testing,** 2. **Angular location of the selected clusters should span the maximum part of the azimuth and zenith domains.**   **Observation 7:** It is not very clear whether the all Set A/B beams can be emulated inside the test chamber or only some of those beams which are “most relevant” Tx-beams can be emulated.  **Proposal 6: TE vendors to clarify the following:**   1. **Whether all Set A/B beams can be emulated inside the test chamber or only some of those beams which are “most relevant” Tx-beams can be emulated. How to define the “most relevant” Tx beams.** 2. **Is it feasible to emulate all Set A/B beams inside the test chamber which arrives at the DUT from different AoAs.**   **Proposal 7: UE vendors to clarify if the Set A/B beams emulated inside the test chamber can’t arrive at the DUT from different AoAs, will it be possible for the DUT to distinguish between the different beams which may have different Beam IDs but will arrive at the DUT from the same AoA.** |
| [**R4-2416477**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416477.zip) | Keysight Technologies UK Ltd | *Observation 1: The AI/ML beam management objective seems primarily targeted for FR2.*  *Observation 2: Three TE vendors provided overall favourable view towards CDL models while making sure that some simplifications are considered.*  *Observation 3: It is advisable to limit the maximum number of spatially separated TRPs to 2 if more than 1 TRP is required*  *Observation 4: 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, set B Tx beams up to [16] with power offsets, and set A Tx beams up to [192] with power offsets.*  *Observation 5: The extension of an FR2 MIMO OTA system to 8 probes is suitable to support unlimited range of AoDs/ZoDs and a range of AoAs/ZoAs of 110°/25°.*  *Observation 6: 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 7: 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.*  *Observation 8: It is not clear whether the benefit of AI/ML UE-side models outweighs the impact in performance, e.g., reduced battery life which is very critical for handhelds (PC3)*  **Proposal 1: It is proposed to limit the testability discussions of AI/ML Beam Management test cases to FR2 only.**  **Proposal 2: Do not consider a single-probe IFF system for AI/ML Beam Management use cases due to lack of CDL support**  **Proposal 3: Do not consider the Enhanced IFF system supporting RRM and Multi-RX test cases for AI/ML Beam Management use cases due to large probe separation and lack of absolute probe locations.**  **Proposal 4: Consider the FR2 MIMO OTA test system (or derivatives thereof) with DFF probes to properly and accurately implement CDL channel model.**  **Proposal 5: UE/chipset/infra vendors to provide feedback on whether modelling 4 RX-PAS accurately with the remaining beams modelled with power offsets is sufficient**  **Proposal 6: Consider an 8-probe OTA system with 2 additional probes when compared to the FR2 MIMO OTA system with a range of AoAs/ZoAs of 110°/25° as the starting point for AI/ML Beam management testing for 1 TRP and/or 2 TRPs with close angular proximity.**  **Proposal 7: Based on the overview of Table 1 and Observation 7, select CDL models for the testability aspects of the AI/ML Beam Management use cases** |
| [**R4-2416493**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416493.zip) | Samsung | *Testing target*  **Proposal 1:** AI-BM testability study shall be performed to test UE-side model inference requirement as 1st priority, including (1) testing the accuracy of inference results; (2) testing the delay of reporting inference results; (3) generalization performance testing if introduced.  **Observation 1:** For UE-sided model, DL Tx beam prediction, testability discussion could be different for AI/ML model input by using the measurements from (1) the best Rx beam or (2) Tx-Rx beam pair.  **Proposal 2:** AI-BM test setup shall generate the not only Set B of beams, but also Set A of beams for the ideal measurement of RSRP on the predicted Tx beam.  *Test procedure for AI-BM (UE-side model)*  **Proposal 3:** The conformance testing procedure for AI-BM UE-side model, using AI-BM Case1 as a reference, is detailed as follows:  Preparation Phase:  - Step-1: TE configures resource sets for both Set A and Set B beams in *CSI-ReportConfig*, and associated ID as side condition  Model Inference Phase:  - Step-2: TE configures the test environment to transmit Set B beams  - Step-3: DUT (a.k.a, UE) predicts Set A beams based on the measurement from Set B beams, and reports the prediction results back to TE  Ideal Measurement Phase:  - Step-4: TE adjusts the test environment to transmit Set A beams, ensuring a certain level of SNR for the ideal measurement of RSRP  - Step-5: DUT measures the Set A beams, and report measurement results to TE  TE Evaluation Phase:  - Step-6: TE evaluates the success of the prediction by comparing the predicted results from Step 3 with the actual measurement results from Step 5.  *Beam prediction KPI*  **Proposal 4:** For metrics for UE-sided AI-BM requirements/tests, RAN4 adopt the combined 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: K = 2 and x = 3 as baseline   *Channel model for UE-sided AI-BM testing*  **Observation 2:** The necessity of introducing spatial differentiated AoAs in AI-BM testing could depends on different AI/ML model inputs:   * For AI/ML model with TX-RX beam pair measurement as model input, CDL model with the spatial differentiated AoAs is required. * For AI/ML models by TX beam measurement from the best RX beam as model input, if an ideal spherical coverage (uniform in all directions) is assumed, to test this AI/ML models, the spatial AoA characteristics could be not that important.   **Proposal 5:** At least for Rel-19 AI-BM, RAN4 discussion shall focus on single TRP case, which is considered as baseline assumption for testability study.  **Proposal 6:** 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:**  - RAN1 SLS assumption can be used as baseline:  🡪 Certain channel model, e.g., UMa with distance-dependent LoS probability function defined in Table 7.4.2-1 in TR 38.901 can be used as baseline.  🡪 FFS UE mobility included in this evaluation scenario  🡪 BS TX Beambook assumed  🡪 For DFF/IFF chamber, FFS how UE RX antenna modeling; For 3D MPAC OTA chamber, RX antenna modeling is precluded.  **Step-2: SLS for channel model for all TX beams:**  - Based on the targeted evaluation scenario, run SLS to get channel model for training/testing  🡪 To obtain the data for different UE location, multiple drops will be simulated in SLS.  🡪 For each drop, the resultant channel models are obtained for all beams in Set-A/Set-B  **Step-3: Test signal is mapped over probe(s) in OTA chamber**  - The signals from tester (to be over certain TX beam-i) + corresponding channel model for TX beam-i (obtained in Step-2), to be mapped over the probe(s) in OTA chamber.  - Depends on DFF or 3D MPAC OTA chamber is decided to be used, the mapping can be different.  **Proposal 7:** RAN4 shall use UMa model as used in RAN1 evaluation to perform SLS for generating the received signals from different TX beams, in order to obtain the dataset for AI model training/evaluation.  **Proposal 8:** For FR2 TX beambook, the reference model used in TR 38.827 can be used as baseline for further discussion:  - For FR2: A code book of 128 fixed beams is constructed to a grid of eight elevation angles from –25° to +25° with ~7.1° step size and 16° azimuth angles from –60° to +60° with 8° step size.  *OTA testing chamber for AI-BM model inference*  **Proposal 9:** The feasibility of enhanced 3D MPAC chamber shall consider the following factors:   1. FFS how many beams out of Set A (e.g., 128 beams containing Set B beams) are relevant for AI-BM inference testing; 2. All Tx beams from a single TRP; 3. Acceptable PAS similarity percentage based on a certain configuration of probes |

## Open issues summary

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

1. Rx beam knowledge
2. TCI state knowledge
3. Absolute RSRP definition
4. Relative RSRP accuracy
5. Requirements for beam prediction
6. Measurement error impact evaluation
7. Beam prediction KPIs
8. Channel model for beam prediction testing
9. Test setup requirements
10. AoA/AoD
11. Beam prediction testing process

### Sub-topic 2-1

*UE Rx beam known vs. unknown*

In the previous meeting it was discussed whether the UE Rx beam to be used for receiving the predicted beam is known or not.

**Issue 2-1: UE Rx beam knowledge**

Proposals

* + Option 1: UE Rx beam is known
    - the UE Rx beam gain has to be taken into account when making the prediction
  + Option 2: UE Rx beam is unknown
  + Option 3: depends on some conditions
    - which conditions?
  + Option 4: others
* Recommended WF
  + Option 1

### Sub-topic 2-2

*TCI State known vs. unknown*

In the previous meeting it was discussed whether the TCI state associated with a predicted beam should be considered known or unknown.

**Issue 2-2: TCI State knowledge**

* Proposals
  + Option 1: TCI state is known if the RS is QCL-ed with an RS measured within 1280ms
  + Option 2: TCI state QCL to an RS that is not in Set B is unknown even if this RS was included in the L1-RSRP report by UE within 1280ms before TCI state activation.
  + Option 3: 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.
  + Option 4: The known conditions of the predicted TCI state shall take below conditions into account:
    - The TCI state switch command is received during inference phase of AI/ML mode.
    - The predicted beam in Set A is the RS in target TCI state.
  + Option 5: RAN4 should consider relaxing known conditions for TCI states
  + Option 6: others
* Recommended WF
  + To be discussed

### Sub-topic 2-3

*RSRP Absolute accuracy*

**Issue 2-3: RSRP Absolute accuracy**

* Proposals
  + Option 1: Absolute RSRP accuracy= predicted L1-RSRP of beam index i – ideal measured L1-RSRP of beam index i. i may be any beam index in top-K predicted beams.
  + Option 2: other
* Recommended WF
  + Option 1

### Sub-topic 2-4

*Relative RSRP accuracy*

**Issue 2-4: Relative RSRP Accuracy**

* Proposals
  + Option 1: The definition of relative RSRP accuracy shall be:
    - 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 value (including predicted and measured beams), if beam index n is not configured for corresponding measurement.
    - Relative RSRP accuracy= predicted L1-RSRP of beam index i – measured L1-RSRP of beam index n, where the beam index n owns the largest value (including predicted and measured beams), if beam index n is configured for corresponding measurement.
  + Option 2: Relative RSRP accuracy is the RSRP difference between the predicted L1-RSRP of different beams
  + Option 3: Relative RSRP accuracy is modified as follows:
    - * **Relative RSRP accuracy = predicted L1-RSRP of beam index i – ~~predicted~~ measured L1-RSRP of predicted beam index n, where the beam index n owns the largest ~~predicted~~ measured value obtained by measuring the predicted beams.**
  + Option 4: ***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***
* Recommended WF
  + Option 2

### Sub-topic 2-5

*Requirements for beam prediction*

**Issue 2-5: Requirements for beam predictions**

* Proposals
  + Option 1: Requirements to be defined:
    - Measurement period for prediction (how long will set B be measured before making a prediction)
    - Prediction accuracy
    - TCI state known/unknown
  + Option 2: other requirements
* Recommended WF
  + At least requirements listed in Option 1 are needed, FFS if other requirements are needed

### Sub-topic 2-6

*Simulations for error impact evaluation*

In the previous RAN4 meeting it was agreed to conduct a study on the impact of measurement errors. More details about the study need to be discussed

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

* Proposals
  + Option 1: Agree on what simulations to perform
    - SLS – which assumptions to use?
    - LLS – which assumptions to use?
  + Option 2: how to model the measurement errors?
    - is there a need to model errors for both training data or only for inference?
  + Option 3: others
* Recommended WF
  + To be discussed how to move forward on this study, what set of assumptions are needed and what are the next steps

### Sub-topic 2-7

*Beam prediction KPIs*

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

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

- Option 1: RSRP accuracy

- Option 2: Beam prediction accuracy

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

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

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

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

-Related measurement accuracy can be considered to determine x

- Option 4: combinations of above options

**Issue 2-7: Metrics/KPIs for beam prediction**

Proposals

* + Option 1: Use Option 1
  + Option 2: Use Option 2
  + Option 3: Use combined Option 2 and 3:
    - 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: K = 2 and x = 3 as baseline
  + Option 4: Neither Option 1, 2, 3 is appropriate, a new metric is needed
  + Option 5: Other
* Recommended WF
  + To be discussed

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

### Sub-topic 2-8

*Channel model for beam prediction testing*

**Issue 2-8: Channel model for beam prediction testing**

* Proposals
  + Option 1: simplified CDL
  + Option 2: TDL
  + Option 3: others
* Recommended WF
  + Option 1

FFS how to simplify the CDL channel models, there are some proposals in this meeting

### Sub-topic 2-9

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 2-9: Test setup requirements**

* 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?
    - UE rotation during the test
    - multiple TRP (or reflections from another direction)
    - AoA – to be discussed later
    - AoD – to be discussed later
  + Option 2:
    - other parameters
* Recommended WF
  + To be discussed

### Sub-topic 2-10

*AoA/AoD*

The test system design will be influenced by the UE AoA range and the TE AoD range

**Issue 2-10: AoA/AoD**

* Proposals
  + Option 1: AoA:
    - range should be large enough to reasonable ensure that UE needs multiple beams to receive the signals within this range
    - other conditions?
  + Option 2: AoD:
    - range should be large enough to ensure enough Tx beams are distinguishable
    - range should be large enough to ensure that enough channel clusters are transmitted
    - other conditions?
* Recommended WF
  + To be discussed

If further study is needed, it should be discussed what analysis is needed

### Sub-topic 2-11

*Beam prediction testing process*

**Issue 2-11: Beam prediction testing process**

* Proposals
  + Option 1: Test needs to have at least 2 parts:
    - 1. inference part of the test (UE has to make the actual prediction)
    - 2. predicted beam measurement (UE has to measure and report “the ground truth”)
  + Option 2: others
* Recommended WF
  + Take Option 1 as baseline, discuss further details of tests such as under what condition should the ground truth be measured, how to ensure the UE does not cheat, etc

# 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** |
| [**R4-2415172**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415172.zip) | CATT | **Proposal 1: Case 3a/3b has no impacts on RRM core requirements.**  **Proposal 2: For case 2b, the existing positioning core requirements can be directly reused.**  **Proposal 3: RAN4 to define a single set of core requirements for case 1 and 2a.**  **Observation 1: RAN2 agreed to use LPP signallings for at least case 1.**  **Proposal 4: RAN4 to reuse the framework of legacy reporting and measurement period requirements. Details are FFS pending more RAN2 inputs.** |
| [**R4-2415192**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415192.zip) | vivo | ***Proposal 1: RAN4 to define delay requirement for case 1 at least considering the following procedures: PRS measuring, AI processing, UE reporting location to LMF.***  ***Proposal 2: RAN4 to consider the straight-line distance (unit: m) between the real position and the estimated position as the metric for accuracy requirement for case 1.***  ***Proposal 3: RAN4 to postpone the discuss of requirement for positioning case 2b.***  ***Proposal 4: RAN4 to define delay requirement for case 2a at least considering the following procedures: PRS measuring, AI processing, UE reporting intermediate features to LMF.***  ***Proposal 5: RAN4 to reuse the existing accuracy requirements for the timing-related intermediate features: RSTD, UE Rx-Tx time difference for positioning case 2a.***  ***Observation 1: It is maybe not possible to define requirement for the LoS/NLoS indicator soft value.***  ***Proposal 6: RAN4 to define performance requirement for LoS/NLoS indicator hard value, the accuracy requirement can be the probability of maximum LoS/NLoS misestimation.***  ***Proposal 7: RAN4 not to define requirements for positioning case 3.***  ***Proposal 8: RAN4 to wait for more progress from other WGs to consider define delay requirements for monitoring for positioning case 1.*** |
| [**R4-2415255**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415255.zip) | CMCC | ***Proposal 1: for case 1, it is proposed to define requirements.***   * ***accuracy requirements: can be defined as the difference in meters/sub-meters between reported position and ground truth labels in each dimension of x-axis and y-axis*** * ***ground truth: according to RAN1 discussion, LMF can derive the ground truth label, e.g. by PRU or target UE and/or gNB sends measurement (e.g., legacy measurement) to LMF so that LMF can derive the information on ground truth label***   ***Proposal 2: for case 2a, it is proposed to define requirements for LOS/NLOS indicator, at least for hard indicator.***  ***Proposal 3: for case 2b, it is proposed to take legacy measurement accuracy requirements for DL RSTD, UE Rx-Tx and PRS-RSRP as baseline.***   * ***According to RAN1 agreements, for AI/ML based positioning case 2b, timing information and the paired timing information and power information are supported for UE reporting to LMF.***   ***Proposal 4: 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, e.g. gNB Rx-Tx time difference absolute accurac and gNB SRS-RSRP, can be conaisdered as baseline.***   * ***According to RAN1 agreements, for AI/ML based positioning case 3b, timing information and the paired timing information and power information are supported for UE reporting to LMF.*** * ***According to RAN1 agreements, for AI/ML based positioning case 3a, at least LOS/NLOS indicator and/or timing information (e.g. UL RTOA or gNB Rx-Tx time difference) are supported for reporting.*** |
| [**R4-2415501**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415501.zip) | Apple | **Proposal 1: RAN4 to further discuss the feasibility and how to define requirements for positioning accuracy for case 1. Defining requirements is also dependent on RAN1/2 conclusion regarding whether UE will report positioning**  **Proposal 2: For Assisted AIML Positioning, the KPI test metric (e.g., LOS/NLOS) needs to be considered for validating the positioning accuracy. RAN4 should inevstiagte how to establish requirements for LOS/NLOS indicators, considering their potential impact on performance and accuracy in positioning systems.**  **Proposal 3: For AIML positioning, RAN4 should take as baseline the requirements on the measurements supported in legacy positioning (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 4: RAN4 to study defining performance accuracy requirements for any reported UE measurements defined by other groups.**  **Proposal 5: RAN4 should define measurement requirements for both path-based and sample-based reporting in relation to use case 2b.**  **Proposal 6: Before RAN4 investigates testing Case 2b in AI/ML for positioning, it is important to first examine the relationship between the measurement accuracy reported by the UE and the resulting positioning accuracy at the LMF.**  **Proposal 7: 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 8: 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** |
| [**R4-2415626**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415626.zip) | Huawei,HiSilicon | **Proposal 1: RAN4 to wait for RAN1 conclusion before discussing new RRM requirements for data collection for AI/ML positioning.**  **Proposal 2: RAN4 to discuss requirements for LCM procedure especially performance monitoring for AI/ML positioning, based on RAN1 conclusion on performance monitoring schemes and also RAN4 conclusion on the requirements for inference.** |
| [**R4-2415892**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415892.zip) | Qualcomm Incorporated | **Observation 1: RAN2 chair has guided companies to focus on 1st priority positioning use cases. RAN plenary has suggested RAN1 not to discuss any proposal targeting 2nd priority issues in Q4 of 2024.**  **Observation 2: UE based positioning with UE side model is a 1st priority item of AI/ML positioning.**  **Observation 3: Legacy spec has previously defined the signalling for reporting UE’s location via DL-TDoA IE in non-AI-ML scenario.**  **Proposal 1: RAN4 postpones defining requirements for LOS/NLOS metric until RAN1 and RAN2 make further progress regarding case 2a, including LOS/NLOS metric.**  **Proposal 2: RAN4 studies defining requirements for Case 1, i.e., UE-based positioning with UE-side model.** |
| [**R4-2416330**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416330.zip) | Ericsson | **Observation 1**: No work is expected to be done by RAN1 on positioning use cases 2a/2b in Q4 2024.  **Proposal 1**: RAN4 to postpone discussions on issues related to requirements for AI/ML positioning use case 2a/2b until Q1 2025.  **Observation 2**: Existing report mapping for UL-RToA can be reused to report measurements for positioning use case 3a.  **Observation 3**: Existing report mapping for SRS-RSRPP can be reused to report path power measurement for positioning use case 3a/3b.  **Proposal 2**: Extend applicability of existing report mappings for UL-RToA and UL SRS-RSRPP to report UL measurements for AI/ML based positioning use cases. |
| [**R4-2416365**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416365.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.***  ***Observation 2: RAN4 would not define any positioning accuracy requirements because positioning is LMF based.***  ***Proposal 1: RAN4 shall not define the positioning accuracy requirements for case 2b, it is up to network implementation.***  ***Observation 3: 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 4: For AI/ML based positioning, the main difference compared to the legacy is that the AI model resides within the LMF.***  ***Observation 5: RAN4 would not define any positioning accuracy requirements since positioning is LMF based. Apart from the positioning accuracy requirements, the reported metrics can be studied for case 2a.***  ***Proposal 3: The intermediate features such as RSTD, RSRP, UE Rx-Tx time diference etc. or some enhancements shall be considered.***  ***Observation 6: 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 when considering PRS measurement and AI model processing time.***  ***Observation 7: RAN1 just agreed at least LOS/NLOS indicator is supported for reporting.***  ***Proposal 5: Although LOS/NLOS could be the intermediate feature to be reported, RAN4 shall consider whether and how to define the requirements for LOS/NLOS.***  ***Proposal 6: RAN4 shall deprioritize the study of ToA until RAN1 supports UE report ToA as model output for case 2a.*** |
| [**R4-2416447**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416447.zip) | Nokia | 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. 2. RAN4 requirements (e.g., in Cases 1, 2a) should be firstly defined with respect to single PFL. 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. 4. RAN4 to study the requirements on a quality indicator for the ground-truth accuracy on positioning label in data collection. 5. Positioning accuracy should be considered as the performance metric in Case 1 (UE-based positioning with UE-side model, direct AI/ML positioning). 6. 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. 7. 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. 8. 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. 9. RAN4 to define the time latency limit on UE’s LCM actions indicated by network in Cases 1 and 2a. 10. 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. 11. 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). 12. For the report mapping of DL PRS-RSRPP (or UL SRS-RSRPP), the existing mapping of measured quantity is defined in Table 10.1.38.3.1-1 (or Table 13.6.1-1) can be reused/used as a baseline. 13. RAN4 to study the reporting requirements of PRS-RSRPP / SRS-RSRPP for AIML positioning (e.g., report mapping, reporting resolution, etc.). |
| [**R4-2415032**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415032.zip) | OPPO | **Proposal 1: For case1, RAN4 will not define positioning accuracy requirements in R19 WI.**  **Observation 1: For case1(AI/ML assisted positioning), the feasibility of using 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.**  **observation 2: For case2a, the feasibility of using intermediate results(output of AI/ML models with non-linear processing) for case 2a 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.** |
| [**R4-2415173**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415173.zip) | CATT | **Proposal 1: Ground truth of DUT can be set in tests as well as the coordinates of TRPs. RAN4 can provide these coordinates by proper designs of test cases.**  **Proposal 2: RAN4 to define relative accuracy requirements based on 3-D distance difference between ground truth and UE predicted position with following side conditions considered:**  **- SINR**  **- SCS and BW of measurement resources, e.g., PRS**  **- Channel model**  **- Other key parameters, if identified.**  **Proposal 3: RAN4 to consider tightening the existing accuracy requirements for AI/ML enabled positioning.**  **Proposal 4: RAN4 not to define requirements for LOS/NLOS indicator reporting.** |
| [**R4-2415193**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415193.zip) | vivo | ***Observation 1: PRUs can be deployed in the test scenario to provide measurements together with the location information.***  ***Proposal 1: RAN4 to consider deploy PRUs in conformance or the pre-deployment test so as to derive the ground truth value of UE position and intermediate features.***  ***Observation 2: At least LoS/NLoS indicator hard value can be tested since the LoS/NLoS environment is known in test system.***  ***Proposal 2: RAN4 to discuss the key factors that affect the performance and consider introduce reference scenario/model/dataset for AI/ML based positioning test for generalization.*** |
| [**R4-2415254**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415254.zip) | CMCC | ***Proposal 1: for case 1, it is proposed to define tests.***  ***Proposal 2: for case 1, in order to get ground truth, test area can be grid consisting of location point(s) with known coodinates. And PRU can also be considered.***  ***Proposal 3: case 2a and case 2b, it is proposed to define tests for reported metrics.*** |
| [**R4-2415502**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415502.zip) | Apple | **Proposal 1: If the report of positioning from UE has been established by other WG, RAN4 to investigate the feasibility and testability for Case 1, where a UE predicts its position relative to ground truth obtained through physically located PRUs with known grid locations or derives the ground truth through GPS.**  **Proposal 2: RAN4 to investigate emulating various propagation conditions like single path channel for Line-of-Sight (LOS) scenarios and multipath channels for Non-Line-of-Sight (NLOS) conditions in the test equipment. Ground truth can be established by legacy LOS/NLOS estimation.**  **Proposal 3: RAN4 can begin investigating the related testability of new defined metrics (like CIR/PDP) once it has received sufficient input from other groups**  **Proposal 4: RAN4 should investigate the relationship between the measurement accuracy provided by the UE and the eventual positioning accuracy at the LMF before establishing any requirements.** |
| [**R4-2415627**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415627.zip) | Huawei,HiSilicon | ***Proposal 1:*** ***For UE-sided direct AI/ML positioning (Case 1), there is no need to define the requirement of positioning accuracy.***  ***Proposal 2:*** ***For UE-assisted/LMF-based positioning with UE-side model (Case 2a), RAN4 will not define requirement of AI-based LoS/NLoS indicator.***  ***Proposal 3:*** ***For UE-assisted/LMF-based positioning with UE-side model (Case 2a), RAN4 will study the testability of AI-based timing information if introduced by other WGs.***  ***Proposal 4:*** ***For UE-assisted/LMF-based positioning with UE-side model (Case 2a), RAN4 to clarify which exact metrics are considered as “existing metrics” and should meet the existing requirements.***  ***Proposal 5:*** ***For UE-assisted/LMF-based positioning with LMF-side model (Case 2b), existing requirements apply for existing metrics as baseline.***  ***Proposal 6:*** ***For UE-assisted/LMF-based positioning with LMF-side model (Case 2b), RAN4 will will continue to postpone the discussion of requirement in Case 2b.***  ***Proposal 7:*** ***For NG-RAN node assisted positioning with gNB-side model (Case 3a) and NG-RAN node assisted positioning with LMF-side model (Case 3b), RAN4 will not define requirement of gNB reporting to LMF.*** |
| [**R4-2416331**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416331.zip) | Ericsson | **Observation 1**: No agreement has been reached in RAN4 to define requirements for case 1.  **Observattion 2**: Requirements to be tested for case 1 need to be understood before starting discussion on testability aspects related to positioning case 1.  **Observation 3**: Progress made by RAN1 on positioning cases 2a and 2b is limited and is not enough to make effective discussion on testability aspects related to these positioning cases in RAN4.  **Proposal 1**: RAN4 to postpone discussion on testability issues related to positioning cases 2a and 2b until sufficient progress that allows RAN4 to trigger discussion on testability issues is made by other working groups on these use cases.  **Proposal 2**: RAN4 to conclude that discussion on the testability aspects related to positioning case 3a/3b is not needed. |
| [**R4-2416364**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416364.zip) | ZTECorporation,Sanechips | **Observation 1: If the ground truth is known to test equipment, the accuracy can be quantified compared to defined ground truth, such as AI/ML positioning error. Then the model performance can be evaluated and the subsequent operation of model can be clarified.**  **Observation 2: The UE location is the coordinate and the PRU will be set on anywhere, if we set PRU randomly, we do not think the accuracy can be accurately evaluated.**  **Observation 3: If UEs have the GNSS capability, they also can be the reference UEs. However, if we choose the reference GNSS UEs, the GNSS error shall be considered when evaluating the AI/ML positioning accuracy error.**  **Proposal 1: The ground truth can be the PRU and the test map shall be considered. Reference GNSS UEs shall be considered.**  **Observation 4: RAN1 deems that LOS/NLOS indicator can be the input for model inference at LMF side. However, how to verify the indicator is one concern since RAN4 did not discuss the propagation conditions before.**  **Proposal 2: Whether the single path or multi channels shall be considered and which kind of channels shall be used shall be studied.**  **Proposal 3: A single path for LOS is enough. How many multiple channels will be used for NLOS shall be discussed based on which kind of channels will be used.** |
| [**R4-2416448**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416448.zip) | Nokia | 1. 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). 2. **RAN4 should specify (i) the test mechanisms for positioning accuracy and (ii) the metric for positioning accuracy performance in Case 1.** 3. Based on the earlier RAN4 #110 agreement not to specify positioning accuracy requirements for Case 3a and Case 3b, there are no testability aspects for Case 3a and Case 3b, except if RAN1 specifies new measurement types for AIML-based positioning or enhances existing measurement types for AIML-based positioning.   **RAN4 to consider testability aspects for Case 3a and Case 3b only, if RAN1 specifies new measurement types for AIML-based positioning or enhances existing measurement types for AIML-based positioning.** |

## Open issues summary

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

1. Requirements for case 1 – positioning accuracy
2. Requirements for case 3
3. Report applicability for existing report mappings
4. Test definition for case 1

### Sub-topic 3-1

*Requirements for case 1*

Several companies proposed to define requirements for case 1

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

* Proposals
  + Option 1: Introduce requirements for case 1
    - requirements:
      * accuracy requirements: can be defined as the difference in meters/sub-meters between reported position and ground truth labels in each dimension of x-axis and y-axis
      * ground truth: according to RAN1 discussion, LMF can derive the ground truth label, e.g. by PRU or target UE and/or gNB sends measurement (e.g., legacy measurement) to LMF so that LMF can derive the information on ground truth label
    - RAN4 to consider the straight-line distance (unit: m) between the real position and the estimated position as the metric for accuracy requirement for case 1
    - relative positioning accuracy metric (change in position due to a change in environment)
    - other requirement/definition
  + Option 2: do not introduce any requirement for case 1
* Recommended WF

### Sub-topic 3-2

*Requirements for case 3/3b*

In the previous meeting it was agreed to wait for further RAN1 agreements to decide whether any requirements will be defined or not

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

* Proposals
  + Option 1: 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, e.g. gNB Rx-Tx time difference absolute accurac and gNB SRS-RSRP, can be conaisdered as baseline.
    - According to RAN1 agreements, for AI/ML based positioning case 3b, timing information and the paired timing information and power information are supported for UE reporting to LMF.
    - According to RAN1 agreements, for AI/ML based positioning case 3a, at least LOS/NLOS indicator and/or timing information (e.g. UL RTOA or gNB Rx-Tx time difference) are supported for reporting
  + Option 2: RAN4 to define performance accuracy requirements for use case 3a/3b for the measurements reported by gNB. The specifics and scope of the core and performance requirements for use case 3a/3b are contingent on the outcomes of discussions within RAN1 and RAN2
  + Option 3:RAN4 will not define any requirement for gNB reports to LMF
  + Option 4: others
* Recommended WF
  + To be discussed

### Sub-topic 3-3

*Reference applicability for existing metrics*

**Issue 3-3: Report applicability for existing reported metrics**

* Proposals
  + Option 1: Extend applicability of existing report mappings for UL-RToA and UL SRS-RSRPP to report UL measurements for AI/ML based positioning use cases.
  + Option 2: others
* Recommended WF
  + Option 1

### Sub-topic 3-4

*Testing for Case 1*

Several companies proposed to define tests for case 1 and also made some proposals on how to determine the ground truth

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

* Proposals
  + Option 1: RAN4 to introduce tests for case 1
    - ground truth can be determined through use of PRUs or GNSS
  + Option 2:RAN4 to postpone testing discussions until the requirement definition is better understood
  + Option 3: RAN4 will not define any tests for case 1
  + Option 4: others
* Recommended WF
  + To be discussed

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

# 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** |
| [**R4-2415028**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415028.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: For RAN4 test model alignment, the motivation should be further clarified**   * **Option1: It is expected that the alignment of minimum performance understanding can be achieved through results from different companies** * **Option2: It is expected to strictly align the output results of each company's AI/ML models through simulation**   **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 above 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: According the refined simulation parameters in R4-2414447, corresponded second-round alignment results could be found as below:**   |  |  |  | | --- | --- | --- | | CSI feedback bits | Quantization of encoder output | SGCS (UMa channel) | | 64 bits | without quantization of the FC layer outputs for validation purposes only, i.e., float32 is sent in CSI report | **0.717** | | 32 floats | 2 bits quantization for CSI report | **0.771** |   **Observation 4: 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**  **Observation 5: For the reference 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.** |
| [**R4-2415072**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415072.zip) | Xiaomi | **Proposal 1: RAN4 should discuss and establish the criteria for simulation alignment.**  **Proposal 2: If alignment cannot be achieved using company-specific data, RAN4 should discuss whether and how to combine data from different companies to create a common data set for training models.**  **Proposal 3: Multiple decoders can be chosen as candidates first and then the final decoder can be selected from these candidates with good generalization.**  **Proposal 4: For data formats, Numpy or Matlab is fine.**  **Proposal 5: For model format, RAN4 to further discuss the feasibility of ONNX.** |
| [**R4-2415174**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415174.zip) | CATT | **Observation 1: SGCS increases from 75% to 78% after simulation assumptions were refined.**  **Proposal 1: RAN4 can randomly select a decoder after performance is well aligned. eType II performance can optionally be taken into account.**  **Observation 2: For 4a, different model backbone can be chosen by UE vendor and TE vendor and these models may perform badly with each other.**  **Proposal 2: RAN4 documents both reference decoder and dataset with potential key aspects specified, such as quantization method.**  **Proposal 3: For Option 4c study, companies to run simulations with their own trained encoder and the chosen decoder and report the results for alignment.**  **Proposal 4: Verifying the TE verification/validation is out of RAN4 scope.** |
| [**R4-2415187**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415187.zip) | vivo | **Proposal 1: 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 2: In Option 4a-1 (the specified dataset is generated by the aligned reference encoder and reference decoder), the model structure needs to be specified, to guarantee the performance of test decoder.**  **Proposal 3: In Option 4a-1 (the specified dataset is generated by the aligned reference encoder and reference decoder), dataset of raw channel or encoder input needs to be aligned, by aggregating the dataset of raw channel or encoder input from all companies, or choosing one dataset of raw channel or encoder input from all companies.**  **Proposal 4: The feasibility of Option 4a-2 (the specified dataset is aggregated by the datasets provided by all companies) could not be concluded, since at least the encoder from aggregated dataset could not work.**  **Proposal 5: In Option 4b (reference encoder based), channel generation method and reference decoder structure need to be in the spec.**  **Proposal 6: There is no essential difference between reference model and reference dataset. Reference dataset may have some disadvantages compared to reference model, and the benefit of reference dataset over reference model is unclear.**   * **Reference dataset would cause some performance loss compared to reference model.** * **The spec impact of writing model and dataset in the spec would be similar.** * **Reference dataset needs extra step compared to reference model, which is dataset generating from the aligned reference encoder and reference decoder.** * **The storage size of reference dataset is larger than reference model.** * **Reference model is more flexible than referend dataset for future release.**   **Proposal 7: Reference encoder and reference decoder are needed for requirement derivation, and could largely be used for test decoder derivation for both Option 3 and Option 4 (Option 4a-1 and Option 4b). RAN4 to work on reference encoder and reference decoder firstly. Later to discuss what will be put in the spec.**  **Proposal 8: Similar feasibility conclusion through the simulation campaign of test options can be draw for both Option 3 and Option 4 (Option 4a-1 and Option 4b).**  **Proposal 9: One test decoder could be used for one test case or multiple test case.**  **Proposal 10: Different reference encoder may be defined for different requirement or test case.**  **Proposal 11: Option 3 next steps:**   1. **Check on performance alignment -> see simulation results from contributing companies**    * **repeat simulations until good alignment is achieved**    * **move to next step after alignment** 2. **Share models (encoder or decoder or both)/datasets (training/testing/inference)**     * **framework to share data among companies(dataset/model) to be discussed** 3. **Select one or more decoder for further analysis**    * **selection criteria: select the decoder with the best performance on the aggregated dataset with only encoder input** 4. **Each company brings results for training of “own encoder” with selected decoder(s)**    * **performance alignment to be checked/discussed**    * **using “own” data or data shared by other companies** 5. **Conclude on overall feasibility of Option 3**    * **consider the conditions under which Option 3 is feasible if found feasible**    * **feasibility criteria: nearly all companies can obtain acceptable performance using their “own encoder” paired with selected decoder(s).**   **Proposal 12: Option 4a-1 next steps:**   1. **Check on performance alignment -> see simulation results from contributing companies**    * **repeat simulations until good alignment is achieved**    * **move to next step after alignment** 2. **Share models (encoder or decoder or both)/datasets (training/testing/inference)**     * **framework to share data among companies(dataset/model) to be discussed** 3. **Select one or more dataset(s) for further analysis**    * **selection criteria: select the dataset that is generated by the model pair with the best performance on the aggregated dataset with only encoder input** 4. **Each company brings results for training of “own encoder and decoder” with selected dataset(s)**    * **performance alignment to be checked/discussed**    * **using “own” training method or training method shared by other companies** 5. **Conclude on overall feasibility of Option 4a-1**    * **consider the conditions under which Option 4a-1 is feasible if found feasible**    * **feasibility criteria: nearly all companies can obtain acceptable performance using their “own encoder and decoder” paired with selected dataset(s).**   **Proposal 13: Option 4b next steps:**   1. **Check on performance alignment -> see simulation results from contributing companies**    * **repeat simulations until good alignment is achieved**    * **move to next step after alignment** 2. **Share models (encoder or decoder or both)/datasets (training/testing/inference)**     * **framework to share data among companies(dataset/model) to be discussed** 3. **Select one or more encoder for further analysis**    * **selection criteria: select the encoder with the best performance on the aggregated dataset with only encoder input** 4. **Each company brings results for training of “own decoder” with selected encoder(s)**    * **performance alignment to be checked/discussed**    * **using “own” or data shared by other companies** **data** 5. **Conclude on overall feasibility of Option 4b**    * **consider the conditions under which Option 4b is feasible if found feasible**    * **feasibility criteria: nearly all companies can obtain acceptable performance using their “own decoder” paired with selected encoder(s)**   **Proposal 14: The feasibility study of Option 3, Option 4a-1 and Option 4b could work in parallel.**  **Observation 1: Based on the aligned model structure and simulation assumptions, the SGCS of AI/ML model for 64 bits payload is 0.75, AI/ML model for 64 float value is 0.784, and the SGCS of eType II is 0.73. The CDF curve of SGCS is shown in Figure 2.3-1.**   |  |  |  |  | | --- | --- | --- | --- | |  | **eType II** | **AI/ML model**  **for 64 bits** | **AI/ML model**  **for 64 float value** | | **SGCS** | **0.720** | **0.732** | **0.784** |   Figure 2.3-1. CDF curve of SGCS for aligned model structure and simulation assumptions  **Observation 2: It is seen that AI/ML model has better mean performance but worse variance performance, compared to eType II. The loss function for training may be further studied in the future.**  **Proposal 15: In future actual reference encoder and reference decoder design, RAN4 may consider smaller model, for the convenience of implementation.**  **Proposal 16: 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 3: 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.** |
| [**R4-2415251**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415251.zip) | CMCC | ***Proposal 1: for the feasibility study of option 3, the evaluation results are shown as following:***  Table SGCS for CSI compression based on CNN model   |  |  | | --- | --- | |  | SGCS | | CNN | 0.708 | |
| [**R4-2415376**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415376.zip) | NTU | **Proposal 1: Revise option 4a and 4b to ensure test repeatability:**  **- 4a: Dataset based: specify at least dataset based on which the test decoder can be implemented by TE vendors**   * **4a-1: the test decoder has to satisfy**     **where is the set of decoder inputs (latent messages) in the specified dataset, and *f* is the chosen loss/similarity function to evaluate the difference between the decoder output from the dataset corresponding to the latent message *z*, , the test decoder output from *z*, , and is a RAN4 specified threshold.**   * **4a-2: the test decoder has to satisfy**   **where the notations are the same as 4a-1, and is the test decoder output space.**  **- 4b: Reference encoder based: encoder is documented in the specifications, used to derive a decoder to be implemented by TE vendors**   * **4b-1: the test decoder has to satisfy**     **where is the reference encoder output (latent message) of encoder input *x*, and is the sampled encoder input in the test decoder verification procedures. Note that RAN4 needs to specified the generation procedures of *X*.**   * **4a-2: the test decoder has to satisfy**   **where is the test decoder output space and *Z* is the reference encoder output space.**  **Observation 1: Option 4a-1 and 4b-1 require both TE vendors and UE vendors to train the test decoder and DUT encoder with the dataset or reference encoder, while option 4a-2 and 4b-2 allow TE vendors and UE vendors to train the test decoder and DUT encoder with arbitrary datasets or encoders/decoders as long as the specified ordering requirements are satisfied.**  **Proposal 2: 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 2: 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 3: Additional margin to account for the mismatch need to be considered when determining requirement.**  **Proposal 4: Consider the test decoder selection criterion based on the agreed step, i.e., check whether performance alignment across contributing companies’ “own encoder”, which may have different structures, can be achieved.  Therefore, companies can bring the test decoder proposal together with performance metrics achieved when connecting to different encoders with different structures, e.g., the structure options for test decoder discussion, but trained with the decoder input and output dataset from the proposed decoder.** |
| [**R4-2415384**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415384.zip) | MediaTek inc. | **Observation #1: For the case of 2GHz and 1 layer, the SGCS of quantized ResCNN model is 0.697 with 64 bits feedback. Also, the SGCS of eTypeII PC1 with 62 bits feedback is 0.719.**  **Observation #2: For the case of 4GHz and 1 layer, the SGCS of quantized ResCNN model is 0.695 with 64 bits feedback. Also, the SGCS of eTypeII PC1 with 62 bits feedback is 0.716.**  **Observation #3: For the case of 2GHz and 4GHz with 1 layer, the comparison of ResCNN and eTypeII CDFs of SGCS indicate that the performance losses surpass the performance gains, leading to an average decrease in overall performance.**  **Observation #4: FLOPs of ResCNN encoder is 93M and FLOPs of ResCNN decoder is 1480M.**  **Observation #5: Parameters of ResCNN encoder is 0.127M and parameters of ResCNN decoder is 1.808M.** |
| [**R4-2415498**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415498.zip) | Apple | **Observation 1: Simulation results comparing SGCS between AI/ML and e-typeII are captured below:**   |  |  | | --- | --- | | **SGCS** | **Encoder/Decoder** | | **0.698 (Relative improvement 1.9%)** | **AI/ML** (64 bits) | | **0.685** | **e-type II** (configuration 1, 62 bits) |   **Observation 2: We observe a tendency for overfitting when we train the AI/ML model. This can be observed by the continuous improvement of the training error and the saturation of the testing error.**  **Observation 3: Edge sub-bands seem to have the worst performance in terms of SGSS for the AI/ML model**  **Observation 4: Significant performance differences are observed (in SGCS) across different simulation assumptions for Option 3 test decoder**  **Proposal 1: Each company to report training/testing results across training epochs. If overfitting is consistently observed, investigate the sources of overfitting**  **Proposal 2: Each company to report cdf plots of the testing data per SGCS. Based on these results investigate if there is a tendency for the edge sub-bands to perform worse.**  **Proposal 3: To better understand the relative gains of AI/ML compared to type-II SGCS in adapting optimized simulation parameters for feasibility analysis, we propose conducting a related study. This will help pave the way for simulation alignment and for the definition of performance requirements. We have identified two parameters for optimization:**   1. **Learning rate (it could be adaptive)** 2. **Phase normalization of the eigenvectors** 3. **Increasing the data size remedies tendency to overfit**   **Proposal 5: RAN4 to employ the procedures shown in the flowchart for studying the feasibility analysis of test decoder Option 3. (It captures both Option 1 and Option 2)**  A diagram of a flowchart  Description automatically generatedv  **Proposal 6: 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 7: 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)**   **FFS on how to define a criterion for the similarity/dissimilarity of all companies' performance results with respect to the reference performance, considering UE encoder complexity as well**  **Proposal 8: 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:**   * **Each company tests its trained pair of {Encoder, test decoder) with other companies testing data set. Chose the pair (and therefore test decoder) that has the best performance averaged across other’s companies datasets.**   **Proposal 9: According to the flowchart, RAN4 should first evaluate feasibility analysis with Option 1 and proceed to Option 2 (aggregated dataset) if Option 1 does not produce a feasible test decoder**  **Proposal 10: RAN4 to employ the procedures shown in the flowchart for studying the feasibility analysis of test decoder Option4 based on the aggregate dataset and database of test decoder models. (Option 5 for partially specified test decoder)**   1. **The architecture of the test decoder is partially specified through some high-level parameters (to be agreed) while leaving room for vendor specific implementations** 2. **Interested companies implement a test decoder according to the partially specified agreement** 3. **The partially specified test decoder is trained from an aggregate data set of UE encoder latent space distributions (dataset is stored in a server). Multiple test decoders are trained from participating companies. The test decoders are stored in a database.** 4. **Each UE vendor tests its own preferred Encoder with the test decoders stored in the database** 5. **Analysis of results to determine feasibility of test decoder**   A diagram of a diagram  Description automatically generated  **Proposal 11:**  **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** |
| [**R4-2415621**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415621.zip) | Huawei,HiSilicon | ***Proposal 1:*** Use throughput as test metric for AIML-enabled CSI compression.  ***Observation 1:*** MLP is superior to CNN in terms of SGCS.  ***Observation 2:*** The complexity of MLP is lower than that of CNN.  ***Proposal 2:*** Principle of reference model design – complexity of reference model is one of the metrics for selecting reference model structure.  ***Proposal 3:*** Different model structures of encoder and decoder can be considered for reference model design, e.g., MLP for encoder while Transformer for decoder.  ***Proposal 4:*** Two options are identified for defining the alignment in step 1.   * Option 1: The SGCS margin is within x   + FFS: the value of x * Option 2: The throughput margin is within y   + FFS: the value of y   ***Proposal 5:*** Take throughput performance gain into account for defining the criteria of overall feasibility for potential testing options.  ***Observation 3:*** The feasibility of option 4 is the same as that of option 3.  ***Proposal 6:*** If the overall feasibility of potential testing options can be verified, RAN4 will further discuss which part of the model would be specified if AI CSI compression turns to WI. |
| [**R4-2415688**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415688.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: Further investigate feasibility of using the following two options for training encoder**   * **Option 1: Encoder training based on the encoder output / decoder input dataset** * **Option 2: Encoder training based on the decoder output dataset**   **Proposal #3: Update the simulation assumptions for Option 3 feasibility analysis based on Table 1.**  **Proposal #4: Randomly select 2-3 candidate decoder implementations for further feasibility analysis.**  **Proposal #5: Further study Option 4 under assumption that the following information is specified**   * **Channel dataset** (encoder input / decoder output) * **Latent space dataset** (encoder output / decoder input) * **Reference encoder/decoder**   **Proposal #6: 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.** |
| [**R4-2415889**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415889.zip) | Qualcomm Incorporated | **Observation 1: The mean SGCS of CNN based model and eTypeII based on latest simulation assumptions are shown below:**   |  |  |  | | --- | --- | --- | | **Method** | **Mean SGCS**  **(across samples and subbands)** | **Median SGCS**  **(across samples and subbands)** | | **CNN encoder – CNN decoder without quantization** | **0.74** | **0.8** | | **CNN encoder – CNN decoder with quantization** | **0.683** | **0.73** | | **eTypeII** | **0.67** | **--** |     **Observation 2: In option 4b, UE vendor needs to know the reference decoder to train its encoder. Hence, for option 4b to work, reference decoder needs to be specified along with the reference encoder.**  **Proposal 1: Option 4b is suggested to be modified in the following way (modifications are shown in blue):**   * **4b: Reference encoder and reference decoder based: encoder is documented in the specifications, used to derive a decoder to be implemented by TE vendors. Reference decoder is documented in the specifications, used by UE to train its encoder.**    + **FFS whether training dataset (channel information) needs to be specified** |
| [**R4-2415965**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415965.zip) | Nokia | **On Results of decoder/encoder training:**  Table 2: Trained AI/ML encoder-decoder pair SGCS performance with 2-bit quantization and 64-bits overhead.   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Dataset** | **Samples** | **Training SGCS** | **Testing SGCS** | **Epoch** | **Model**  **Parameters** | **Rel16 SGCS** | **SGCS gain Rel 16** | | **NOKIA** | **630K** | **0.721** | **0.705** | **99** | 1.93M | **0.691** | **2.0%** | | OPPO | 600K | 0.728 | 0.703 | 138 | - | - | | CAICT | 100K | 0.715 | 0.676 | 75 | - | - | | Mixed | 300K | 0.710 | 0.686 | 73 | - | - |  1. RAN4 to consider the Table 2 above to collect SGCS performance for CSI encoder-decoder pairs trained on diverse datasets with two-bit quantization of CSI feedback.   Table 3: SGCS of Reference CNN model without Quantization.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Dataset** | **Samples** | **Training SGCS** | **Testing SGCS** | **Epoch** | **Model**  **Parameters** | | **NOKIA** | **630K** | **0.769** | **0.755** | **150** | 1.93M | | OPPO | 600K | 0.775 | 0.752 | 150 | | CAICT | 100K | 0.766 | 0.729 | 80 | | Mixed | 300K | 0.761 | 0.718 | 143 |  1. RAN4 to consider the Table 3 above to collect SGCS performance for CSI encoder-decoder pairs trained on diverse datasets without quantization of CSI feedback.   Table 4: Cross-validation of encoder-decoder par with different datasets.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Train on** | | **Test on** | | | | | **NOKIA** | **OPPO** | **CAICT** | **Mixed** | | **NOKIA** | **630K** | 0.705 | 0.645 | 0.595 | 0.648 | | **OPPO** | **600K** | 0.626 | 0.707 | 0.639 | 0.657 | | **CAICT** | **100K** | 0.497 | 0.586 | 0.677 | 0.587 | | **Mixed** | **300K** | 0.680 | 0.689 | 0.688 | 0.686 |  1. As shown in Table 4 above, the trained reference encoder-decoder pairs exhibited dataset dependency, indicating that a model trained on a single dataset may not perform well on other datasets, even under identical scenario settings. Training on a mixed dataset significantly enhances the model’s generalization ability, with larger mixed datasets yielding better performance.   **On derivation of test decoder and encoder with Option 3:**  Additional criteria for the selection of decoder(s) from the set contributed by all the companies are needed. The decoder demonstrating best performance on one dataset may not be that efficient on another dataset, e.g., in the channel emulator implementation provided by the TE vendor.   1. RAN4 can consider for the further study three decoders: with highest, lowest and closest to mean relative performance gain in SGCS over eTypeII decoder.   Use of aggregated dataset {*V*} of channel eigenvectors can be used for much more objective performance comparison/validation of encoders/decoders contributed by the companies.   1. RAN4 to use an aggregated dataset of 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. When an aggregated dataset is used for model training, a single decoder demonstrating best SGCS performance can be selected for the further study.   Similarly to the selection of the test decode, the performance of the encoder trained based on given decoder can demonstrate difference performance on different datasets/channel generators from different vendors.   1. To conclude about the feasibility of Option 3, RAN4 needs to ensure that there is no significant performance variation in between “own” encoder – test decoder(s) by different companies in different datasets/aggregated dataset. 2. RAN4 would need to check the span of the throughput-based test metric in between the companies to conclude about the success of Option3 feasibility study.   **On Partially specified decoder with Option 4:**  The specified dataset for Sub-option 4a should consist of channel/eigenvector and latent message pairs {*V, c*}. The latent messages *c* in the dataset for Sub-option 4a are readily obtained from a reference encoder. Test decoder performance specifications for verification for Sub-option 4a can be set based on a testing portion of the specified dataset.  Datasets for training the test decoder in Sub-option 4b can be created by the test equipment vendor using the reference encoder. A standardized testing dataset can be used in Sub-option 4b as the basis for performance verification of the test decoders.  Sub-options 4a and 4b differ in how the training dataset is obtained. For consistency of test decoder performance between test equipment vendors, a common specified verification dataset is desirable.   1. For further analysis of Sub-option 4c, additional clarification of constraints “on the decoder output ordering” is requested of its proponents. 2. RAN4 to adopt the following procedure for the feasibility study steps of Sub-option 4a (specified dataset): 3. **Generate a dataset of channel/eigenvector realizations {*V*}** 4. **Create reference encoder/decoder.** 5. **Generate dataset {*V, c*} – standardized, common dataset.** 6. **Companies design their own test decoders (replicating the design process of TE vendors) using their architecture of choice, training and verifying them using the standardized {*V, c*}.** 7. **Companies train their own encoder (replicating the UE vendor design process) using an architecture of their choice.**    1. **For training their encoders, companies could choose between two alternatives: Alternative 1: Use the standardized dataset Alternative 2: Use their own dataset** 8. **Perform testing of all the UE encoders using all test decoders (replicating the process of testing UE’s against RAN4 requirements with different TE vendor decoders).** 9. RAN4 to adopt the following procedure for the feasibility study steps of Sub-option 4b (reference encoder): 10. **Generate a dataset of channel/eigenvector realizations {*V*}.** 11. **Create reference encoder/decoder.** 12. **Create a verification test dataset (common).** 13. **Each company (mimicking a test equipment vendor) generates their own training dataset {*V, c*} using the common reference encoder to create the encoder output *c* from the encoder input *V*.** 14. **Companies design their own test decoders (replicating the design process of TE vendors) using their architecture of choice, training them using the dataset {*V, c*} from the previous step and verifying them using the verification dataset from Step 3.** 15. **Companies train their own encoder (mimicking a UE vendor) using an architecture of their choice, training the encoder using a dataset different from the dataset in the previous step.** 16. **Perform testing of all the UE encoders using all test decoders (replicating the process of testing UE’s against RAN4 requirements with different TE vendor decoders).**   **On Verification of test decoder:**   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. |
| [**R4-2416314**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416314.zip) | Ericsson | [Observation 1 A sufficiently well trained test decoder and sufficiently rich range of test conditions is needed to ensure that the test decoder approach provides basic interoperability.](#_Toc179214563)  [Observation 2 The reference encoder is useful for training network side decoders and for validating TE implementations.](#_Toc179214564)  [Observation 3 For option 4, it is essential to ensure that the information is sufficient that all TE vendors get to the same ordering of the latent space representation](#_Toc179214565)  [Observation 4 It is important that monitoring can observe and quantify partial reduction in performance (not just a large or catastrophic loss).](#_Toc179214566)  [Observation 5 It is non-trivial to ensure that a monitoring function can determine the performance of an AI model in isolation, since in the real world a very large number of interacting factors influence performance.](#_Toc179214567)  [Observation 6 For the 2-sided model, a further challenge with monitoring is allocating observed performance degradation to the two half models.](#_Toc179214568)  [Observation 7 Monitoring requirements should define the constancy/accuracy of monitoring reports from the UE if they are defined.](#_Toc179214569)  [Observation 8 Testability of monitoring report accuracy needs discussion, since the reliability of underlying models will differ between UEs and a test of accuracy needs to force UEs into known reliability levels.](#_Toc179214570)  [Observation 9 The current RAN4 study has not identified the best training set or decoder structure, since the aim is to investigate the ability to converge to a common model, not to select the right model.](#_Toc179214571)  Based on the discussion in the previous sections we propose the following:  [Proposal 1 The RAN4 test decoder in option 3 should be trained with a sufficiently wide and rich dataset that it can provide good performance over the range of normal operating conditions (i.e. channel type, Doppler etc.)](#_Toc179214575)  [Proposal 2 For option 3, also provide a reference encoder.](#_Toc179214576)  [Proposal 3 Discuss means to ensure convergence in the latent space representation for differently trained decoders for option 4. Standardization of a reference decoder is one potential option.](#_Toc179214577)  [Proposal 4 As well as mean SGCS, SGCS distributions should be compared.](#_Toc179214578)  [Proposal 5 Compare the following between companies to obtain insight into the degree of alignment of the training/testsets: Amplitude CDF, strongest Eigenvalue CDF, SGCS for 2 and/or 4 strongest beams.](#_Toc179214579) |
| [**R4-2416349**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416349.zip) | ZTE Corporation, Sanechips | ***Observation 1. The results of quantization non-aware training is superior to quantization aware training.***  图形  Fig 2. CDF curves for CSI compression (Quantization aware training)  图形_Quantization_non-aware_training  Fig 3. CDF curves for CSI compression (Quantization non-aware training)  ***Observation 2. The FLOPS of the CNN model is significantly higher than other models, e.g. transformer, MLP.***  ***Observation 3. The performance of the CNN model is inferior to that of other models, e.g. transformer, MLP.***  ***Observation 4. RAN4 needs to build a strict guidelines to govern process for Release 19 for AI compression test metric.***  ***Observation 5. If RAN4 would like to study feasibility for CSI compression, new regulations should be considered for AI/ML-based KPI.***  ***Proposal 1. RAN4 may need to investigate how large a span can be considered aligned results.***  ***Proposal 2. To consider reference encoder based method, and training datasets need to be specified.*** |
| [**R4-2416395**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416395.zip) | Rohde & Schwarz | **Observation 1: For the CNN-based (Resnet) model selected for the Option 3 feasibility study, we observe a SGCS of 0.7728 after considering the alignment decisions described in R4-2414447. The deviation from our previously reported reconstruction accuracy results is marginal.**  **Observation 2: The computational complexity of the CNN-based (Resnet) test decoder remains unchanged in comparison to our previous evaluation (1,475 MFLOPs).** |
| [**R4-2416495**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416495.zip) | Samsung | *Standardization for Test decoder Option 3*  **Proposal 1:** For this round of performance alignment, at least the following issues shall be studied and concluded:   * Issue-1: Whether the performance alignment can be achieved, with the aligned CNN-structure and training procedure (both of which are aligned as much as possible), by leaving the following difference:   + the company-wise SLS-generated dataset, with particular SLS implementation, randomness and the size of dataset. * Issue-2: How performance results can be regarded as being “aligned”?   **Proposal 2:** On the issue of “*How performance results can be regarded as being ‘aligned’*”, the following steps are required to be studied:   * Alignment Step-a: On the evaluation dataset provided by individual company, what is the SGCS performance spread among companies?   If the alignment is not confirmed in Alignment Step-a, consider Alignment Step-b:   * Alignment Step-b: On a common evaluation dataset (agreed to be used by the group), what is the SGCS performance spread among companies?   **Proposal 3:** 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).  **Observation 1:** SGCS performance and complexity for CNN-based model trained and evaluated over Samsung-SLS generated dataset is provided as:   |  |  |  | | --- | --- | --- | | 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 | 0.63 |   **Observation 2:** An increased number of UEs per drop leads to a greater dispersion in channel characteristics, which in turn can adversely affect the SGCS performance (Note: this is observed across all models evaluated on the Samsung dataset.)  **Observation 3:** The SGCS performance varies significantly across different SLS simulated datasets, which utilize varying numbers of UEs per drop.  **Observation 4:** The importance of common dataset for training and/or evaluation has been demonstrated.  *Discussion on Test decoder Option 4*  **Observation 5:** The feasibility of Option 4b depends upon the confirmed feasibility of Option 3.  **Observation 6:** For Option 4a, if the training data set (including enough amount of data for raw CSI and compressed bit strings) is specified in 3GPP standard, Option 4a can be regarded as the standardized training dataset.  **Proposal 4:** Among three sub-options, Option 4a is preferred by assuming:  - TE vendor will not share decoder directly to other vendors (DUT and/or infra vendors);  - Parameters that need to be specified for defining test decoder shall include:  🡺 Training data set for TE decoder training, including enough amount of data for raw CSI and compressed bit string. |
| [**R4-2414959**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2414959.zip) | CAICT | **Proposal 1: Suggest to consider TDL channel as starting point for CSI prediction tests. FFS whether to introduce CDL channel.**  **Proposal 2: For performance test of CSI prediction, suggest to take static condition as default assumption.**  ***Observation 1:*** *Two possible alternations for introducing non-static conditions in generalization test:*   * *Alt 1: the conditions are randomly changed* * *Alt 2: the conditions are sequentially changed in predefined order.*   **Proposal 3: Further discuss whether and how to introduce non-static conditions in generalization test.**  ***Observation 2:*** *The definition of ground truth for CSI prediction is analogous to beam management.*  **Proposal 4: The discussion on ground truth shall wait for RAN1’ s progress on input and output of CSI prediction.**  **Proposal 5: Suggest to follow legacy testing as starting point, i.e., random precoding use TypeI single panel codebook as baseline for relative throughput.**  **Proposal 6: Further discuss other alternatives, for example, random precoding using the same codebook as prediction as the baseline for relative throughput.**  ***Observation 3:*** *For AI-based CSI prediction, the CSI information predicted based on measurements within one measurement period or one measurement window is able to cover multiple reporting instances.*  **Proposal 7: Suggest to discuss taking UE reported precoder generated from measurement as baseline for relative throughput. FFS the specific test procedure (e.g., as illustrated in figure 1 and 2).** |
| [**R4-2415175**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415175.zip) | CATT | **Proposal 1: RAN4 to use non-AI based PMI as the baseline of accuracy metric.** |
| [**R4-2415189**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415189.zip) | Vivo | **Observation 1: In non-AI CSI prediction, random precoding based on Type I Single Panel codebook has been agreed as the baseline throughput**  **Proposal 1: The baseline of AI CSI prediction could be the throughput from random precoding, which means that AI prediction and non-AI prediction could follow the similar test procedure.**  **Proposal 2: RAN4 to discuss simulation assumptions and reference model for AI CSI prediction.** |
| [**R4-2415250**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415250.zip) | CMCC | ***Proposal 1: for CSI prediction accuracy metrics for inference, the relative throughput is the throughput gain achieved with predicted PMI compared to random PMI.***  ***Proposal 2: for CSI prediction, it is proposed to use intermediate KPI, e.g. SGCS, as requirements/tests metrics for LCM.*** |
| [**R4-2415358**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415358.zip) | SEU | **Observation 1:** ***Problems regarding the similarity between the training and testing datasets have not been discussed.***  **Observation 2:** ***The metric for monitoring the model input data has not been fully discussed.***  **Observation 3:** ***The trade-off between the accuracy performance and generalization has not been fully considered.***  **Observation 4:** ***The generalization ability of cell-specific framework should be evaluated.***  **Proposal 1: *The evaluation metrics regarding the performance of performance monitoring should also be considered.***  **Proposal 2: *True positive rate and true negative rate can be used as metrics to evaluate the validity of performance monitoring methods.***  **Proposal 3: *The ability of identifying unqualified prediction outcomes should be given greater emphasis in performance monitoring.***  **Proposal 4: *The concept of generalization must be expanded to encompass not only spatial variations across different scenarios but also temporal dynamics within those scenarios.***  **Proposal 5: *Both in-distribution and out-of-distribution cases should be evaluated for performance monitoring.*** |
| [**R4-2415386**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415386.zip) | MediaTek inc. | **Observation #1: Existing PMI reporting tests use random PMI of Type I codebook as reference throughput.**  **Proposal #1: Use random PMI of Type I codebook as reference throughput in test metric for CSI prediction.**  **Observation #2: When defining Release 18 PMI prediction test requirements it was challenging to find test configuration that enables measurable PMI prediction gains over legacy baseline.**  **Proposal #2: We propose to use eType II non-predicting reporting as reference in test case definition to ensure test feasibility.** |
| [**R4-2415499**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415499.zip) | Apple | **Observation 1: In the context of CSI prediction performance monitoring, RAN1 is already exploring a network-side mechanism that relies on UE-reported ground truth. This same approach can be utilized to evaluate the accuracy of CSI prediction as a performance KPI within RAN4**  **Proposal 1:** **The γ value comparing predicted PMI to random PMI can be used in the test for CSI prediction. To infer the gains of AI/ML-based CSI prediction, the γ values can be compared between AI/ML and non-AI/ML cases.**  **Proposal 2: Same mechanism used for NW-side performance monitoring can be employed for testing where the UE provides both the current channel information and the predicted channel information. Subsequently, the TE can calculate the SGCS by comparing the current channel from the latest report with the predicted channel from the previous report.** |
| [**R4-2415623**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415623.zip) | Huawei,HiSilicon | ***Proposal 1***: The baseline of the requirement in CSI prediction is the throughput that TE uses a random PMI.  ***Observation 1***: How to ensure that the testing dataset aligns well with training dataset is still an open issue.  ***Proposal 2***: RAN4 will start to discuss the test configuration after other WGs achieve sufficient progress. |
| [**R4-2415871**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415871.zip) | Ericsson | [Observation 1 The goal with the RAN4 requirement is not to demonstrate AI performance relative to performance with non-AI (“show a gain”), but it is to establish a minimum performance level for the AI/ML-based CSI (PMI) prediction.](#_Toc179226609)  [Observation 2 It is important that monitoring can observe and quantify partial reduction in performance (not just a large or catastrophic loss).](#_Toc179226610)  [Observation 3 It is non-trivial to ensure that a monitoring function can determine the performance of an AI model in isolation, since in the real world a very large number of interacting factors influence performance.](#_Toc179226611)  [Observation 4 Monitoring requirements should define the constancy/accuracy of monitoring reports from the UE if they are defined.](#_Toc179226612)  [Observation 5 Testability of monitoring report accuracy needs discussion, since the reliability of underlying models will differ between UEs and a test of accuracy needs to force UEs into known reliability levels.](#_Toc179226613)  [Observation 6 For network sided monitoring, if the UE reports both predicted and measured CSI, there may be some need to consider setting conformance requirements on the performance of both measured and predicted when operating in this monitoring mode.](#_Toc179226614)  Based on the discussion in the previous sections we propose the following:  [Proposal 1 For the performance metric, use throughput with AI predicted CSI (PMI) relative to random PMI with e.g., Rel-15 Type I Single-Panel Codebook.](#_Toc179226615)  [Proposal 2 Study the variation of performance of CSI prediction with different channel model, SNR level, changes in SNR and Doppler with well trained models and with over-fitted models.](#_Toc179226616)  [Proposal 3 Depending on the results of the study, define sufficient requirement variations that proper model training (i.e., a model/functionality that can perform well across variations in cell conditions) is achieved.](#_Toc179226617)  [Proposal 4 Consider whether to make conformance to CSI prediction with lower/higher Doppler scenario, according to RAN1 study.](#_Toc179226618) |
| [**R4-2415890**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415890.zip) | Qualcomm Incorporated | **Observation 1: PMI prediction related test cases of Rel-18, e.g., “Multiple PMI with 16Tx Enhanced Type II codebook for predicted PMI” (6.3.2.1.8 of [2]) use relative throughput as requirement metric. Random precoding-based throughput is used as the baseline in those test cases.**  **Proposal 1: RAN4 selects random precoding-based throughput as the baseline in the test cases for AI-ML based CSI prediction where relative throughput is used as requirement metric.**  **Proposal 2: The test case for AI-ML based CSI prediction can follow the flow of PMI prediction related test cases of Rel-18.** |
| [**R4-2415967**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415967.zip) | Nokia | **On eType2Doppler-r18 CSI/PMI reporting:**  **Observation 1**: CSI prediction feature and corresponding etypeII-doppler-r18 codebook type was introduced in Rel-18 NR\_MIMO\_evo\_DL\_UL WI, including reporting of multiple predictions (eType2DopplerN4-r18 >1).  **Observation 2**: The RAN4 Demod requirements on Multiple PMI with 16Tx Enhanced Type II codebook for predicted PMI are defined for typeII-doppler-r18 codebook type, N4=1 and for TDLA30-20 channel conditions. The requirements are specified on (γ) - the throughput relative to the random precoding from typeI-SinglePanel codebook.  **On Rel-19 AIML-based CSI reporting:**  **Observation 3**: RAN1 considered UE-side model only and assumed R18 eType II doppler codebook for CSI report for both AI/ML and Non AI/ML predictions.  **Proposal 1: RAN4 to assume R18 eType II Doppler codebook for CSI report for AI/ML prediction. Correspondingly, R16 eType II codebook is supported by the UE as well.**  **Proposal 2: RAN4 can start the discussion with a single AIML-based CSI prediction (N4=1).**  **Observation 4**: RAN1 baseline scenario and model training uses UMa channel model according to TR 38.901.  **Proposal 3: RAN4 to assume that the training of AI/ML CSI prediction model is done based on system-level scenario (e.g., based on UMa channel model according to TR 38.901).**  **Observation 5**: Due to unrealistic spatial behavior of TDL models, CSI prediction would have only trivial outcomes: spatial characteristics of the channel either demonstrate always the same behavior for medium or high correlation, or random for - low correlation.  **Proposal 4: RAN4 to consider Spatially Correlated Channel models (SCM) for the requirements/test on AIML CSI prediction.**  **On metric for RAN4 requirements:**  **Observation 6**: RAN1 has discussed two metrics (SGCS and user TPut) and two benchmarks (the nearest historical CSI and non-AI/ML based CSI prediction) for the evaluation of the gain of AIML -based CSI prediction. Additionally, performance monitoring, including a possibility of ground-truth reporting (Type 2), is under discussion but still lacking the details.  **Observation 7**: RAN4 does not need to use the metric from RAN1 that is based on closest historical CSI for the definition of the reference user throughput because just latest reported CSI can be used.  **Proposal 5: RAN4 can consider further two possible baselines to compare AIML-enabled predicted CSI:**   1. **Use throughput based on randomized typeI-SinglePanel codebook like in legacy test** 2. **Use throughput based on eTypeII codebook measured and reported by the UE (without prediction)**   **On Non-static scenarios:**  **Observation 8**: Non-static scenarios in RAN4 need to be considered 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.  **Proposal 6: RAN4 needs to define a test when some of conditions (e.g., channel type, SINR, speed/Doppler, etc.) change during the test without the change of the functionality.**  **On Generalization, test coverage, and LCM:**  **Proposal 7: RAN4 should discuss the parameters in Table 4 to define the generalization scenarios for CSI prediction use case.**  **Observation 9**: Less generalized functionalities across a set of scenarios can result in frequent switching of model/functionality resulting in performance degradation.  **Proposal 8: RAN4 needs to design a new metric, indicative of generalization capabilities of AI/ML functionality, to verify it’s the generalization performance in different scenarios.**  **Observation 10:** 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.  **Proposal 9: 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 prediction functionality.**  **Proposal 10: If localized models are considered for CSI prediction, RAN4 will need to study model ID-based LCM requirements for a given functionality.** |
| [**R4-2416351**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416351.zip) | ZTE Corporation, Sanechips | ***Observation 1. There is no difference in testing framework between legacy and AI/ML prediction.***  ***Proposal 1. To consider the reported PMI as the baseline throughput for CSI prediction accuracy metric, directly using the legacy predicted PMI or indirectly using random PMI while simultaneously ensuring that the value γ should exceed the legacy predicted requirements.***  ***Proposal 2. For AI/ML based CSI prediction, the testing framework can reuse legacy testing procedures.*** |
| [**R4-2416494**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416494.zip) | Samsung | **Observation 1:** For relative throughput as requirement metric for CSI prediction CSI prediction accuracy metric for inference, the definition of baseline throughput (as denominator):   * Option 1: Throughput achieved by following random PMI (with the same codebook used for the reported predicted PMI) * Option 2: Throughput achieved by following UE’s last reported PMI given by UE measurement   **Proposal 1:** For relative throughput used as CSI prediction accuracy metric for inference, Option 2 (Throughput achieved by following UE’s last reported PMI given by UE measurement) is adopted for baseline throughput. |

## Open issues summary

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

1. Option 3 simulation results for 2nd round
2. Option 3 for 2-sided model
3. Reference encoder and/or dataset for Option 3, test decoder verification
4. Option 4a next steps
5. Option 4b next steps
6. Option 4c next steps
7. CSI prediction tests

### Sub-topic 4-1

*Option 3 simulation results for 2nd round*

**Issue 4-1: Simulation results discussion**

* Proposals
  + Option 1: See simulation results in the table below.

Table to be included, see xls

* Recommended WF

Discuss simulation results, need for further alignment or other issues to be studied until the next meeting

### Sub-topic 4-2

*Selection criteria for decoder(s)*

It should be discussed how to choose which decoder(s) will be used for the next steps, other companies to train an encoder with these decoder(s)

**Issue 4-2: Selection criteria for decoder(s)**

* Proposals
  + Option 1: Select a single decoder, the decoder with best SGCS
  + Option 2: Select 3 decoders: with highest, lowest and closest to mean relative performance gain in SGCS over eTypeII decoder.
  + Option 3: Select 3 decoders randomly
  + Option 4: others
* Recommended WF
  + To be discussed

### Sub-topic 4-3

*Reference encoder and/or training dataset for Option 3*

Several companies brought up the need to document a reference encoder and/or training data set for Option 3. This would be used for performance evaluation/verification or TE verification

**Issue 4-3: Reference encoder and/or training dataset**

* Proposals for parameters that should be agreed upon:
  + Option 1: A reference encoder should be documented (e.g. in a TR)
  + Option 2: A training dataset should be documented (e.g. in a TR)
  + Option 3: A test dataset should be documented (e.g in a TR)
  + Option 4: others
* Recommended WF
  + To be discussed

Option 2 should be discussed separately

### Sub-topic 4-4

*Option 4a for 2-sided model*

Option 4 was further split into 4 sub-options in the previous meeting in R4-2414323. The next steps on feasibility study for these options should be discussed

**Issue 4-4: Option 4a(Dataset based) for 2-sided model**

* Proposals
  + Option 1: Next steps:
    - 1. Discuss/decide how to generate the training dataset and other assumptions
* need to align the ML model, etc
  + - 1. Check on performance alignment see simulation results from contributing companies
* repeat simulations until good alignment is achieved
* move to next step after alignment
  + - 1. Share models (encoder and/or decoder)/datasets (training/testing/inference) (sharing framework is currently under discussion)
      2. Select one or more datasets for further analysis
* Selection criteria is TBD
  + - 1. Companies bring results for “own encoder and decoder” with selected dataset
* performance alignment to be checked
  + - 1. Conclude Option 4a feasibility
* feasibility criteria to be discussed
  + Option 2: Others
* Recommended WF
  + To be discussed

The steps should be discussed and if any changes and/or additions are needed

### Sub-topic 4-5

*Option 4b for 2-sided model*

The next steps on feasibility study for option4b should be discussed, these are likely very similar to option 4a

**Issue 4-5: Option 4b(Reference encoder based) for 2-sided model**

* Proposals
  + Option 1: Next steps:
    - 1. Discuss/decide how to generate the training dataset and other assumptions, including the encoder model
* need to align the decoder assumptions, etc
  + - 1. Check on performance alignment see simulation results from contributing companies
* repeat simulations until good alignment is achieved
* move to next step after alignment
  + - 1. Share models (encoder and/or decoder)/datasets (training/testing/inference) (sharing framework is currently under discussion)
      2. Select one or more encoder for further analysis
* Selection criteria is TBD
  + - 1. Companies bring results for “own encoder and decoder” with selected encoder
* performance alignment to be checked
  + - 1. Conclude Option 4b feasibility
* feasibility criteria to be discussed
  + Option 2: Others
* Recommended WF
  + To be discussed

The steps should be discussed and if any changes and/or additions are needed

### Sub-topic 4-6

*Option 4c for 2-sided model*

The next steps on feasibility study for option4c should be discussed

**Issue 4-6: Option 4c(decoder output ordering) for 2-sided model**

* Proposals
  + Option 1:
* Recommended WF
  + To be discussed

The steps should be discussed and if any changes and/or additions are needed

### Sub-topic 4-7

*CSI prediction tests*

Several companies brought proposals with more details on how to conduct the CSI prediction tests

**Issue 4-7: CSI prediction tests**

* Proposals
  + Option 1: Use TDL model for testing and static condition (similar to legacy test)
  + Option 2: Define test requirements relative to performance with random PMI Type codebook
  + Option 3: others
* Recommended WF
  + To be discussed

Test environment and conditions to be discussed.

# Topic #5: Logistical issues for CSI compression

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| [**R4-2415029**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415029.zip) | OPPO | **Proposal 1: Regarding training dataset size**   * **Training data size ~1GB up to 10GB(depends on what data will be used and the number of samples) /company (one instance)** * **Total space needed would depend on the number of contributing companies and meetings** * **Upper bound should be ~60GB per company per year, possible ~1TB per year to accommodate all companies**   + **Data sharing might not be needed in every meeting**   + **FFS whether the training data for different test settings or side conditions can be shared and reused**   + **FFS the number of datasets needed for different test settings or side conditions**   **Proposal 2:** **The space prepared for the training set now is enough to cover the test and validation sets**  **Proposal 3:** **Use .npy (Numpy) format to do the data transfer as the starting point**  **Proposal 4: Regarding model size**   * **Model size: ~7MB – 100MB per model** * **Total space needed would depend on the number of companies contributing and meetings** * **Upper bound should be 600MB per company per year -> ~10GB per year**   + **FFS whether a model for different test settings or side conditions could be reused**   + **FFS the number of models needed for different test settings or side conditions**   **Proposal 5:** **Use TensorFlow or PyTorch format to do the model transfer as the starting point**  **Proposal 6: The submission methods and numbering problems(e.g., naming or assigning ID) involved in RAN4's data/model sharing should be decoupled from the data/model IDs discussed in RAN1/RAN2.** |
| [**R4-2415188**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415188.zip) | Vivo | **Proposal 1: For simplicity, it is suggested that the testing data could be directly divided by ordering of samples. For example, the first 10%~12.5% or last 10%~12.5% samples are testing data, and the rest samples of dataset are training data (including validation data).**  **Proposal 2: RAN4 to use Numpy for training/testing/inference data sharing, since Matlab is not open source and would cause legal concerns.**  **Proposal 3: RAN4 to use ONNX for model sharing, since Tensorflow or PyTorch could not be supported by all companies.**  **Proposal 4: The aligned dataset and model would need to be stored for a very long time, since they may be captured in WF, TR or spec. Other datasets or models, which are outcomes of intermediate steps in the RAN4 study, would only need to be stored for several years.** |
| [**R4-2415252**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415252.zip) | CMCC | ***Proposal 1: it is proposed to discuss the dataset file descriptions, e.g. scenario/configurations, so that companies can have common understanding and the shared dataset can be used across different companies.***  ***Proposal 2: for the feasibility study for testing of CSI compression based on CNN model, at least real/imaginary parts of the complex numbers, number of sub-bands, number of TX ports need to be provided in the dataset file descriptions.***   * ***An example of detaset file description is proposed as following:*** * ***The number of TX ports is P. N samples are provided in total. Each sample include M complex numbers corrsponding to a eigenvector concatenated by K sub-bands: W=[w1, w2,..., wK] . wq is the eigenvector for the q-th sub-band channel, 1<=q<=K. wq =[Re{wq,1}, Im{wq,1}, ..., Re{wq,P}, Im{wq,P}].*** |
| [**R4-2415385**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415385.zip) | MediaTek inc. | **Observation #1: Dataset file size in NumPy file format for currently agreed training and validation dataset size in total is approximately 4.4GB.**  **Proposal #1: We propose to reserve enough storage margin for increased dataset sizes.**  **Observation #2: AI model file size in Tensorflow file format for currently agreed ResCNN parameters is approximately 25MB.**  **Proposal #2: We propose to reserve enough storage margin for increased model sizes.**  **Observation #3: Training, validation, and testing datasets can be combined in one dataset.**  **Proposal #3: We don’t need to consider the size of training, validation or testing data sets separately for the data storage requirements. These sets can all be part of one common dataset when stored. The exact split into training, validation and testing data sets from the combined dataset can be further discussed.**  **Observation #4: Python and NumPy are open source and free to use and very widely used.**  **Proposal #4: We propose to use NumPy file format in 3GPP work.**  **Observation #5: We have observed some compatibility issues between different tool versions.**  **Proposal #5: We propose to agree some specific tool versions to be used in 3GPP work.**  **Proposal #6: We propose to use TensorFlow format for AI models in 3GPP work, and we are fine to study ONNX if needed.**  **Proposal #7: We propose all shared content need either proper documentation or explicit definition by 3GPP.** |
| [**R4-2415500**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415500.zip) | Apple | **Observation 1:** **While we agree that the assumptions outlined above are reasonable for the ongoing CSI feasibility study, there may be additional storage requirements due to the following factors:**   * **There is a possibility of increased dataset sizes if we consider diverse datasets across generalization simulation scenarios (speed, SNR, propagation model) as well as scalability scenarios (rank, number of ports, etc). We may need to come up with all possible generalization/scalability simulation scenarios we envision to consider for future study.** * **The optimal dataset size to optimize performance and generalization (less overfitting) has not been understood so far** * **For future CSI use cases, like the space-frequency-time CSI compression even larger dataset sizes will be needed** * **Investigation for more layers. Currently the study focuses on single layer. Memory requirements will increase when we study CSI compression for many layers**   **Observation 2: For the generalization study, varying simulation assumptions and parameters may lead to an increase in the number of models required for the CSI use case.**  **Observation 3: We should limit the discussion on the training dataset size. Training data set size will include data used for training, validation, and testing. Partitioning could be agreed as a separate discussion. Validation set could be separate from testing dataset.**  **Observation 4: We support using Matlab '-v7.3' or Numpy file formats. It is essential to reach a consensus on whether to store the raw channel matrix H or the eigenvectors. If one company pre-processes the raw channel into eigenvectors in a specific way and another company uses a different approach, this could lead to inconsistencies (one company trained a model with one method and performs inference with another’s company dataset). Furthermore, storing only the eigenvectors would prevent comparisons with e-type-II**  **Observation 5: TensorFlow and PyTorch are the two most prevalent formats. If the majority of participating companies are already using PyTorch, it would make sense to consider adopting it for simplicity and consistency across collaborations. Standardizing on PyTorch could streamline workflows and reduce complexity, but it's important to ensure that all stakeholders are aligned before making a final decision. If Pytorch gets selected as one of the AI/ML model format, RAN4 should down-select to .pt files of Pytorch. We also support ONNX format.** |
| [**R4-2415622**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415622.zip) | Huawei,HiSilicon | ***Proposal 1:*** The data format and model format discussed in RAN4 will have no impact on other WGs.  ***Proposal 2:*** For the format of data, there are two options to move forward.   * Option1: The content is aligned. The format is up to company’s choice but needs to be reported. * Option2: Both the content and format are aligned.   ***Proposal 3:*** Follow the minimum principle if RAN4 needs to align the data/model format.  ***Observation 1:*** There may exist regulations issues in some areas to upload a large size of dataset to a server deployed overseas.  ***Proposal 4:*** RAN4 and MCC will discuss the regulations issues for data sharing. |
| [**R4-2415689**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415689.zip) | Intel Corporation | **Proposal #1: Adopt the following Training/Inference dataset formats.**   * **Either NumPy or Matlab (v7.3+) file formats** * **Use a single file for training and validation datasets, and a separate file for inference dataset** * **Use separate dataset files to represent different system-level scenarios and for different number of target MIMO layers.** * **Dataset file includes a single float32 array of size [Ns = 600000, NCH = 2, NSB = 13, NANT = 32]** * **The dataset includes eigenvectors calculated based on average covariance channel matrix for the given MIMO layer for each sub-band based on pre-processing assumptions in WF R4-2414447**   **Select datasets provided by several companies and perform training separately on the respective datasets**  **Proposal #2: AI-ML models are shared in both 1) ONNX format, and 2) PyTorch (or TensorFlow) formats** |
| [**R4-2415891**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415891.zip) | Qualcomm Incorporated | **Observation 1: Matlab is not an open-source software.**  **Observation 2: NumPy is frequently used to store training/testing/inference data in machine learning community.**  **Observation 3: ONNX is an open-source software. On the other hand, PyTorch or TensorFlow based model will allow companies to easily replicate another companies’ model during training and testing.**  **Proposal 1: NumPy can be used to store training/testing/inference data.**  **Proposal 2: Companies should share:**   * + **Dataset containing encoder input per subband**   + **Encoder and decoder model.**   + **An additional file containing the vector quantization or scalar quantization codebook that would be applied to quantize encoder’s output.**   **Proposal 3: RAN4 should specify different contents of dataset along with their dimension in the following way:**   * **N (samples) X nSB (number of subbands) X nLayers (number of layers) X nPorts (number of CSI-RS ports) X 2 (IQ)** * **In other words, dataset should content following info:**   + **1st dimension: Number of samples**   + **2nd dimension: Number of sub-bands**   + **3rd dimension: Number of layers**   + **4th dimension: Number of antenna ports**   + **5th dimension: Real and imaginary**   + **Note: Each element of the dataset will be a floating-point real number**   **Proposal 4: Each dataset file’s name should include:**   * **a unique identifier for the company,** * **meeting number** * **part number**   + **dataset can be split among different parts where each part contains data of a subset of samples.** * **One of the following two options can be selected:**   + **Option 1 (Different files for training and testing)**     - **Dataset file’s name should include info regarding the use case of the dataset (training vs testing)**     - **Note1: An example filename can be “EncoderInput\_Train\_Company1\_113\_Part0”**   + **Option 2 (Same file for training and testing)**     - **RAN4 needs to specify how to split the samples between training and testing**       * **e.g., the 1st 80% can be used for training and the last 20% can be used for testing**     - **Note2: An example filename can be “EncoderInput \_Company1\_113\_Part0”**   + **In note1 and note2, #1 denotes the unique identifier of a company and #113 denotes meeting number.**   **Proposal 5: Companies should share AI/ML model, at least, in ONNX. Companies should be encouraged to share AI/ML model either in PyTorch or TensorFlow, too, so that one company can easily replicate another company’s model if they are using the same format.**  **Proposal 6: For ONNX based AI/ML model, following info should be shared:**   * **Version of ONNX and the corresponding version of OPSET [3]** * **Shape of ONNX input and output.** * **Name of different layers of the model along with ONNX format**   + **Naming convention of layer needs to be agreed.**   + **Note: For example, the bias of the 1st layer of the encoder can be referred to as enc.1.bias.** |
| [**R4-2415895**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415895.zip) | Qualcomm Incorporated | Summary of e-mail discussion on data sharing for AI/ML  **2.1 Data size**  2.1.1 Training data:   * Companies agree that the current assumptions and estimates for the CSI compression use case are reasonable (Training data size ~1GB up to 10GB(depends on what data will be used and the number of samples) /company (one instance) * These assumptions are specific to the CSI compression use case in RAN4 and for the current simulation assumptions. * Larger datasets could be needed/used in future studies depending on the use case. It would be good to maintain some flexibility * Importance of ensuring proper data upload/download speeds and supporting breakpoint resume due to potential network interruptions – to be discussed with MCC   2.1.2 AI/ML Model Storage/Sharing:   * Companies agree that the current assumptions and estimates for the CSI compression use case are reasonable (Model size: ~7MB – 100MB, Upper bound should be 600MB per company per year -> ~10GB per year) * These assumptions are specific to the CSI compression use case study in RAN4 * The upper bound is assuming that a company contributes in each meeting within a year (6 WG meetings)   2.1.3 Testing and inference data:   * Most companies commented that testing and inference data should be a subset of the training dataset so there would be no need to consider the data size separately * Testing data set should be 10-25% of the training dataset (most companies showed 10-12.5%) * Other details on partitioning or how to use the datasets would have to be discussed separately   **2.2 Data Formats:**  2.2.1 Training/Testing/Inference data:   * Possible formats to be used for the dataset which brought up are: NumPy, Matlab and Parquet * Formats will have to be discussed with MCC to see which ones can be used in 3GPP. * Most companies prefer NumPy, some where ok also with Matlab * The details of the data formatting (how data is represented, what is contained in each field, etc) will have to be discussed and agreed after the file format will be decided   2.2.2 AI/ML models:   * Possible formats to be used for the AI/ML models are ONNX, PyTorch and TensorFlow * Formats will have to be discussed with MCC to see which ones can be used in 3GPP * There are pros/cons for each format, most companies seem to favor ONNX as it is open source, however, more discussion is needed to agree on a format * Further details on the format (contents of the files and what information needs to be provided) will have to discussed/agreed after deciding on the format   **2.3 Other Issues**   * How to document the file formats will have to be further discussed   + Some level of documentation will be needed such that there is a common reference and anyone can process the data * Different use cases will likely have different needs and all formats discussed here apply for the CSI compression use case * If there will be a need to specify datasets or models, the level of documentation required might be different(more details might be needed, etc) * The need for an API for data sharing needs to be discussed * The “lifetime” of the data needs to be further discussed, some data used in intermediate steps in the RAN4 study might not have to be stored for a long time |
| [**R4-2415966**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2415966.zip) | Nokia | **On Input to MCC:**  **Proposal 1: There is no need to identify a specific dataset type (e.g. training vs testing) when providing RAN4 input to the MCC.**  **Proposal 2: The upper bound of dataset size for currently considered use-case (CSI compression testing feasibility study in RAN4) is up to ~10GB per company per meeting and up to ~160GB for all contributing companies per meeting.**   * **Larger datasets could be used in the future studies depending on the use case so some flexibility in the dataset size should be allowed.**   **Proposal 3: The upper bound of model size for currently considered use-case (CSI compression testing feasibility study in RAN4) is up to ~100MB per company per meeting and up to ~1.6GB for all contributing companies per meeting.**  **Proposal 4: The retention period of the models/datasets shared only for the study purposes can be the length of the 3GPP release, i.e., 18-24 month.**  **Proposal 5: The retention policy of the models/datasets agreed to be captured in the outcome of the study or as a part of the TR can be the same as for regular TDocs/TRs.**  **Proposal 6: The upper estimate of the storage needed to capture the outcomes of the study (i.e., with long retention policy) can be on the level ~160 GB for datasets and ~1.6 GB for models, i.e., as per one meeting for all companies.**  **Proposal 7: Due to the large size of the files to be exchanged (especially for the datasets) the sufficient upload/download speeds need to be ensured and Large File Support (LFS) for the FTP server would be beneficial.**  **Proposal 8: File(s) belonging to a single model/dataset to be packed into a single zip archive.**  **Proposal 9: It is necessary to agree on the rules for naming/identification of shared model/dataset (zip) files: if unique ID cannot be assigned automatically, file naming template can be agreed in RAN4, e.g., Meeting#\_AI\_DataType\_Company\_Version.zip.**  **On RAN4-specific technical aspects:**  **Observation 1**: Based on our experiment it is possible to train the encoder based on the decoder shared and imported via ONNX, at lest for the latest version of ONNX (1.18.0) and in PyTorch.  **Proposal 10: RAN4 to consider using the latest version (e.g., 1.18.0 and later) of the ONNX (Open Neural Network Exchange) to share the models in between the companies.**  **Proposal 11: It is necessary to include a description of the shared model file(s) (i.e., into the zip archive), specifying at least the content (e.g., encoder/decoder or both) and the inputs/outputs of the model(s) (or the reference to the TDoc with the description of those).**  **Proposal 12: RAN4 to consider sharing dataset in NymPy format (i.e., .npy files). If this is not acceptable to all companies, then open-source Parquet file format can be considered as well.**  **Proposal 13: RAN4 to agree on shared dataset structure for CSI compression use-case based on the encoder input, i.e., specify the dimensions of the dataset (a, b, c, d, e), where, a is the total number of sample; b = 2 – real and imaginary part; c – number of sub-band (13); d – number of ports (32); e – optionally, number of layers (1 or 2).**  **Proposal 14: Metadata/description of the dataset should be provided together with it, i.e., included in the zip archive.** |
| [**R4-2416315**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416315.zip) | Ericsson | **Proposal 1 Strive to further understand the relative costs of different retention policies**  **Proposal 2 Consider retention in the light of the following uses for dataset/models, including whether each usage type should have a different retention policy:**   * **Datasets or models shared between companies on a meeting to meeting basis as part of developing a study or work item** * **Datasets or models submitted as reference models** * **Datasets or models that are formally agreed to be references for deriving requirements** * **Datasets or models that are formally captured as part of a TS**   **Proposal 3 Current agreements on model format should be limited in scope to what is needed in the short term for the RAN4 study. Long term ways to capture model formats should be discussed separately (TBC whether RAN4 or RAN1).**  **Proposal 4 ONNX appears to be the most suitable for short term model-sharing, but some experiments to ensure that it can be made to work may be needed.** |
| [**R4-2416350**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416350.zip) | ZTE Corporation, Sanechips | ***Observation 1. For training data, a substantial amount of storage space is required, which poses a challenge for uploading, storing and downloading.***  ***Observation 2. ONNX still have some significant disadvantages, which include interoperability limitations, performance issues, update synchronization and resource consumption.***  ***Observation 3. Python is capable of reading .mat files and .npy files.***  ***Observation 4. The datasets and models may be subject to regulatory constraints in different regions during the upload and utilization process.***  ***Observation 5. 3GPP may need to further discuss how to ensure the security of the datasets or models and compliance with regulatory constraints in different regions.***  ***Proposal 1. RAN4 at least assess the feasibility of ONNX during the WI phase to ensure that conversions between different framework platforms can be fully and accurately realized.***  ***Proposal 2. All uploaded data needs to have the same pre-processing procedure or detailed description about pre-processing and follow the assumptions of RAN4.*** |
| [**R4-2416496**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_112bis/Docs/R4-2416496.zip) | Samsung | *Sharing data formats*  **Proposal 1:** All the proposed data formats, i.e., .npy (NumPy), .mat (Matlab) and Parquet file format, shall be allowed for AI/ML data sharing.  **Proposal 2:** A meta file shall be provided accompanying with the shared data file, by comprising the detailed data format, e.g., the meaning of each field contained each data sample.  *AI/ML model formats*  **Proposal 3:** All the proposed AI/ML model formats, i.e., ONNX, PyTorch and TensorFlow, shall be allowed for AI/ML model sharing.  *Rule of “lifetime” for data/model sharing*  **Proposal 4:** For different categories of lifetimes for data/model sharing, RAN4 can further study based on the following ones as baseline:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | **AI/ML Model** | | | **AI/ML Dataset** | | | | **Cat-1** | **Cat-2** | **Cat-3** | **Cat-a** | **Cat-b** | **Cat-c** | | **Lifetime  (i.e., minimum storage time in 3GPP server)** | 3 months (i.e., to be deleted 3 months after model submission) | 1 release (i.e., to be deleted after the current release ASN.1 is frozen) | Forever (i.e., to be kept in 3GPP server forever) | 3 months (i.e., to be deleted 3 months after model submission) | 1 release (i.e., to be deleted after the current release ASN.1 is frozen) | Forever (i.e., to be kept in 3GPP server forever) | | **Typical sharing purpose** | e.g., temporarily shared model for cross-company evaluation purpose etc. | e.g., reference model (e.g., reference encoder) used to derive RAN4 requirements | e.g., Standardized reference model (e.g., test model used in RAN4 spec.) | e.g., temporarily shared dataset for cross-company evaluation purpose etc. | e.g., the common dataset to derive RAN4 requirement | FFS the necessity |   **Proposal 5:** The folder (to collect shared AI/ML model and/or dataset) shall be set up, by marking different categories of model/dataset, based on the RAN4-agreed rule of “lifetime” for data/model sharing, e.g.,   * Folder: RAN4\_113\_AI\_CSI\_compression\_CNN\_model\_sharing (Cat-1)   *Rule of file/folder naming and storage*  **Observation 1:** There is no strong necessity observed to align file naming rule for model/data files.  **Proposal 6:** The RAN4 moderator could setup the necessary folder to collect data/model, under which different companies’ sub-folder can be set up, such as:   * Folder: RAN4\_113\_AI\_CSI\_compression\_CNN\_model\_sharing (Cat-1)   + Sub-folder: Company-A, Company-B, ... |

## Open issues summary

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

AI/ML model format

Dataset format

Retention policies

### Sub-topic 5-1

*AI/ML model format*

**Issue 5-1: AI/ML format**

* Proposals
  + Option 1:
  + Option 2:
* Recommended WF

To be discussed

### Sub-topic 5-2

*Dataset format*

**Issue 5-2: Dataset format**

* Proposals
  + Option 1:
  + Option 2:
* Recommended WF

To be discussed

### Sub-topic 5-3

*Retention policies*

**Issue 5-3: Retention policies**

* Proposals
  + Option 1:
  + Option 2:
* Recommended WF

To be discussed