**3GPP TSG-RAN WG1 Meeting #110bis R4-2405288**

**China, Changsha, April 15th – 19th , 2024**

**Agenda item:** 9.11.5

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [110bis][135] NR\_AIML\_air

**Document for:** Information

# Introduction

This is the summary thread for issues related to NR AI/ML study in RAN4. The Rel-19 WID was updated in RP-240774.

# Topic #1: General aspects

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2404281**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404281.zip) | InterDigital Communications | * **Option 1** (model testing for conformance to minimal performance requirement) is necessary for any AI/ML function that is either mandatory for a given feature (i.e., cannot be disabled) OR does not have a non-AI/ML fallback option.
* **Option 2** (testing conformance of performance monitoring) requires that there is at least one fallback option e.g., either a pre-deployment tested model (with option 1) OR there is a non-AI/ML fallback option if the functionality is mandatory for a given feature.

***Observation 1:*** *Model identification is required with Option 2 with* *pre-deployment tested model. The indication of an active model is important for a monitoring function.****Observation 2:*** *A fallback model or the baseline that is tested pre-deployment, meets minimum performance requirement is always required by Option2.*Moreover, we believe that the LCM function shall be able to perform the monitoring and model identification, thus the AIML model management.***Observation 3:*** *A fallback can be either fulfilled by one model meeting conformance using Option 1, by a non-AI/ML option or simply disabling the function (if applicable).****Observation 4:*** *The RSRP accuracy requirements are dependent on a minimum Ȇ/Iot level as a side condition.****Observation 5:*** *The AIML predictions accuracy can be corelated with UE measurement accuracy side conditions.****Observation 6:*** *AIML models should be tested for conditions for which they are applicable in field network deployments.****Proposal 1:*** *Support Option 2 with mandatory fallback with pre-deployment conformance certification.****Proposal 2:*** *Model monitoring function with Option 2 support model update/switching/transfer under gNB control.****Proposal 3:*** *The RSRP requirements for baseline test cases to be maintained for beam prediction and data collection.****Proposal 4:*** *RAN4 agree to develop testing requirements for AIML monitoring and reporting functions.****Proposal 5:*** *Testing requirements for AIML models (option 1 or option 2) are based on a model’s applicability conditions for the deployments in field networks.* |
| [**R4-2404426**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404426.zip) | CATT | **Proposal 1: RAN4 identify the procedures for which the latency requirements need to be defined. The following list can be a starting point:** **- Data collections between different entities;** **- Model/functionality activation/deactivation/switch/selection/fallback/inference (signalling based);** **- Model update / parameters transfer;** **- Model transfer / model delivery (if supported);** **- Other procedures.** **Proposal 2: The details of latency requirement for each procedure should be discussed one by one after RAN1/2 makes enough progress on indication/reporting signalling and mechanism designs which help determine the starting point and end point.****Proposal 3: It is not feasible to validate the performance of AI/ML model/functionality impacted by all updates before they are deployed, considering the workload, expense, impacts on application of AI/ML capability.** **Proposal 4: Post deployment performance shall be guaranteed by the vendors and LCM procedures. Meanwhile, RAN4 will design some test cases for LCM procedures to ensure the minimum performance of the deployed AI/ML model/ functionality, which can be discussed in details in performance part.** |
| [**R4-2404478**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404478.zip) | CAICT | **Proposal 1: Suggest to focus on option 2 for handling post deployment testing.** **Proposal 2: Suggest to identify a general scope of impact on RRM core requirements brought by AI/ML-based BM and positioning. Our initial views are provided in Table 1 and 2 (Appendix).**See table in Sub-topic 1-10 |
| [**R4-2404620**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404620.zip) | NTT DOCOMO, INC. | **Proposal 1: Conducting the conformance testing for AI model/functionality before deployment is not feasible because it may need the large number of tests to satisfy the test coverage and it is not realistic.****Proposal 2: The purpose of performance monitoring should be confirmation whether the AI model/functionality currently in use can still be used or not. On the other hand, the purpose of post-deployment validation should be determination whether the AI model/functionality can be re-activated or not.****Proposal 3: RAN4 should consider the requirements of activation for monitoring purpose as post-deployment validation e.g., the number of trials, metric for re-activation which can be different from requirements before deployment.** |
| [**R4-2404719**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404719.zip) | CMCC | ***Proposal 1: it is benifit to have test to guarantee the AI performance after deployment considering that the UE may employ a new model or untested model.******Proposal 2: for UE-side models and/or UE-part of two-sided models, it is proposed that the scenarios and/or configurations used for generalization can be decided based on the supported configuration reported by UE.******Proposal 3: for generalization, it is proposed to take the requirements for inference as the minimum level performance for generlazation.***  |
| [**R4-2404931**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404931.zip) | Intel Corporation | **Proposal #1: UEs supporting AI/ML features shall be mandated to meet the existing legacy performance requirements with configured/enabled AI/ML functionality for all existing legacy test cases.****Proposal #2: The minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature shall be tested for a static and/or non-static scenario and configuration:*** + - **Static scenario/configuration term means that at least channel model type and SNR settings are fixed and do not change over the test, while specific channel realizations may be dynamic.**
		- **Static scenarios/configurations can be applicable to all use cases.**
		- **Non-static scenarios/configuration can be further considered in application to CSI and beam management temporal prediction use cases. The details of models are FFS and may include non-stationary SNR and other conditions.**

**Proposal #3: Adopt the following framework for post-deployment model (feature/functionality) verification at least for the case when model updates or changes are non-transparent to the network:*** + - **At least some default AI/ML model (feature or functionality) needs to pass conformance testing and be present in the device.**
		- **Any changes or updates to the ML-enabled functionality or feature, which come from UE-side OTT server, shall be tested by the device vendor against RAN4 requirements before the deployment to the UE is performed. Other model updates may be tested by the device vendor against RAN4 requirements.**
		- **The information on whether AI/ML model update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**

**Proposal #4: Define reference AI/ML models for one-sided and two-sided models (both encoder/decoder) for performance requirements definition.**  |
| [**R4-2404945**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404945.zip) | vivo | **Proposal 1: Further study necessity of post deployment testing. For CSI compression, several validation methods can be considered as options for post deployment testing:*** **Method 1: The UE obtains a test dataset containing only the channel (encoder input), obtains the PMI by encoder inference, and then reports the PMI to the NW. NW decide the results.**
* **Method 2: UE obtains the test dataset containing both the channel and PMI.** **UE decide the results.**
* **Method 3: UE reports the channel and PMI to NW. NW decide the results.**

**Proposal 2: On post-deployment test for AI/ML based positioning, label-based model output validation can be used for post-deployment test since the ground truth value of PRU can be obtained by solutions.****Proposal 3: On post-deployment test for AI/ML based positioning, label-based model input validation can be used for post-deployment test by preparing additional reference dataset as the ground truth for model input.****Proposal 4: To ensure the AI performance after device deployment, discuss the following options further*** + **Option 1: Conduct the conformance testing for AI model/functionality before deployment**
		- **FFS on the feasibility**
	+ **Option 2: Design the test to verify the performance monitoring**
		- **Depend on the other WG progress**
		- **Monitoring can be used for managing fallback/model update/model switching/model transfer, if applicable**
	+ **Option 3: Define the test to verify the performance validation together with model transfer/update**
		- **validation scheme can be designed by RAN1**

**NOTE: Option 1, Option 2 and Option 3 are all needed for ensuring the AI performance for different life cycle** |
| [**R4-2405148**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405148.zip) | Huawei,HiSilicon | ***Proposal 1***: If new measurements related to performance evaluation is specified, study the feasibility, testability and necessity of defining the relevant latency requirement. ***Observation 1***: Latency of decision-making at UE is not testable.***Observation 2***: Functionality/model selection and activation share similar latency/interruption if any.***Observation 3***: Functionality/model switching and fallback share similar latency/interruption if any.***Proposal 2***: Latency/interruption requirement is not applicable for functionality/model deactivation. ***Observation 4***: Legacy requirements for existing use in RAN4 may not be applicable when defining AI/ML performance requirements, if the effect of operations from the opposite side is not eliminated or not well controlled.***Proposal 3***: Take functionality-based LCM as the starting point for RAN4 discussion.***Observation 5***: Identified scenarios and/or configurations can be initially interpreted as the scenarios and/or configurations that UE reports by capability signaling.***Observation 6***: A large range of various UE capabilities may be involved, which is problematic for RAN4 to identify a typical configuration/scenario for specifying the test cases.***Proposal 4***: RAN4 will discuss how to specify the identified scenarios and/or configurations per use case in future release, if other WGs can specify the granularity and the capability signaling.***Proposal 5***: According to TR 38.843, the identified scenarios and/or configurations can initially be interpreted as the scenarios and/or configurations that UE report by capability signaling. ***Observation 7***: If legacy test metrics are not valid/testable when defining AI/ML-specific requirements, legacy performance requirements for non-AI cannot be reused.***Proposal 6***: RAN4 will study the minimum level performance, per use case, for identified scenarios and/or configurations (if specified). ***Proposal 7***: Other scenarios and/or configurations are interpreted as the scenarios and/or configurations that are not reported by UE capability for an AI/ML-specific (enhanced) feature.***Observation 8***: There is no need to introduce AI/ML-related requirements in the other scenarios and/or configurations.***Proposal 8***: Performance for other scenarios and/or configurations can be ensured by RAN4 legacy test. ***Observation 9***: There is no benefit identified by introducing channel condition changes during test.***Proposal 9***: Non-static scenarios/conditions and propagation conditions are precluded for defining RAN4 test. ***Proposal 10:*** To ensure the AI performance after device deployment, taking the following option as baseline.* + Option 1: Conduct the conformance testing for ~~AI model/~~functionality before deployment
 |
| [**R4-2405184**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405184.zip) | OPPO | **Proposal 1: For both CSI compression and CSI prediction, existed RAN4 test examples for “reporting of PMI” can be reused or serve as a reference. For BM and positioning, we need to consider whether/how to establish new requirements.****Proposal 2: For CSI, reuse the legacy PMI requirement (compared to random precoding) as a baseline test, and let other options/proposals with higher/generalized performance requirements be further studied.****Proposal 3:** **Stability of the performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes, includes: (1) obtaining a consistent monitoring result by considering multiple evaluating samples within an evaluation window, (2) assessing whether model monitoring should be handled at the UE level or the cell level.****Proposal 4: In R19 RAN4 tests, we should focus on verifying whether the solution under test can work in given scenarios or conditions, and whether it can meet the requirements of RAN4 tests within these scenarios and conditions.****Proposal 5: In R19 RAN4 tests, static test scenarios and configurations should be considered first. After having feasible testing cases for static configurations, then further consider whether to introduce non-static testing scenarios and configurations.****Proposal 6: Regarding the AI/ML capabilities, following aspects should be considered** **- Definition of basic AI/ML capability and corresponding testing requirements** **- Definition of different AI/ML capability levels and corresponding different testing requirements** **- Dynamic AI/ML capabilities****Proposal 7: FFS how to define the Post-deployment validation, FFS the necessity and feasibility, what it means to achieve Post-deployment validation, and [possible test methods]** |
| [**R4-2405190**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405190.zip) | ZTE Corporation | **Proposal 1: The scenario of UE capability report, which is AI authentication, it can be identified. Otherwise, the remaining scenarios can be regarded as the other scenarios for test.****Proposal 2: The minimum level of performance shall be studied per use case.****Proposal 3: The legacy performance under different use cases can be as the baseline to judge the minimum performance of AI-based method for each identified scenario.****Proposal 4: The margin shall be added based on the known performance gain for identified scenario to judge what is the significant degradation for other scenarios and try to guarantee that it will not occur**.**Observation 1: 1. The different use case has the different synthetic channels. 2. One synthetic channels shall be used for the individual use case****Proposal 5: RAN4 shall confirm that the different use case has the different synthetic channels.****Proposal 6: RAN shall consider how to design the synthetic testing data.** |
| [**R4-2405610**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405610.zip) | Ericsson | 1. **RAN4 starts the LCM core requirements discussion focusing on the procedures from the common set which are relevant for both functionality- and model-based LCM, so that whichever option is selected in the end, this initial work will still be relevant.**
2. **The initial RAN4 discussion on the LCM core requirements can include: the requirements scope, type of the requirements, and parameters in the requirements, with the understanding that the actual values in the requirements may differ for the functionality-based and model-based LCM.**
3. **RAN4 to discuss and define upon the need activation delay requirements for LCM.**
4. **RAN4 to discuss and define upon the need requirements for switching between the AI mode and the non-AI mode.**
5. **RAN4 discusses and defines upon the need LCM core requirements for performance monitoring.**
6. **As a part of LCM core requirements for performance monitoring, at least requirements related to monitoring data collection are considered.**
7. **RAN4 to discuss whether and how the LCM core requirements related to performance monitoring differ for the considered use cases.**

**Proposal 8. At least for monitoring data collection, RAN4 will assume RRC\_CONNECTED state for both data generation and reporting.**[Observation 1 For option 1 (conformance test before model update deployment), the verification may need to be done on representative hardware for each device variant.](#_Toc163474832)[Observation 2 For option 1, there may be logistical challenges to establishing that all devices in the field are operating models that have passed compliance testing with relevant hardware.](#_Toc163474833)[Observation 3 For option 2 (robust monitoring), the reporting must enable monitoring of the specific AI functionality under question.](#_Toc163474834)[Observation 4 For option 2, RAN4 requirements may be needed to ensure reporting consistency, quality and timing.](#_Toc163474835) Proposal 9. As a further option relating to post deployment testing, consider the possibility of capturing model input during testing for later testing of new models. * + - Option 3: Capture model input during conformance testing for later testing of new models.

 Proposal 10. For the data capture option, consider whether the captured data needs to be held completely by the UE vendor. |
| [**R4-2405651**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405651.zip) | Samsung | *Testing goals for testing framework/procedure***Proposal 1: Provide the following text proposal to Option 1 of testing goal:** * Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.
	+ ~~FFS~~ how to define the specific AI/ML model is provided as use-case specific manner (e.g., a model captured in RAN4 spec as baseline, or a reference model structure agreed for performance alignment)
	+ ~~FFS~~ how to define that the model is properly conducted is provided as use-case specific manner (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)

*Relation to legacy requirements***Observation 1: Based on existing agreement, for the case with the existing legacy performance, the AI/ML enhanced performance shall be defined by using legacy requirement as baseline.** **Proposal 2: No further study or TR refinement is needed for the general aspect of “relation to legacy requirements”, and detailed requirement can be discussed based on specific use case.** *Static/non-static scenarios/conditions for testing***Proposal 3: For how to define a static scenario/configuration:** **(1) the scenario/configuration shall be determined by “the specific configuration/conditions” associated with the relevant “UE capability of an AI/ML-enabled Feature/FG” under testing;** **(2) the static scenario/configuration determined in (1) shall be maintained unchanged during the test.****Proposal 4: Provide the following text proposal to Option 2 of testing goal:** * Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.
	+ ~~FFS~~ how to define a static scenario/configuration is provided as use-case specific manner (e.g., by defining a related testing dataset based on channel models in TR 38.901)
		- The scenario/configuration shall be determined by “the specific configuration/conditions” associated with the relevant “UE capability of an AI/ML-enabled Feature/FG” under testing
		- The static scenario/configuration shall be maintained unchanged during the test
	+ ~~FFS whether and how to define non-static specific scenarios/configurations~~

*Post-deployment handling***Observation 2: 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.** **Proposal 5: 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.**  |
| [**R4-2405675**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405675.zip) | MediaTek inc. |  **Observation 1**: If DUT switched/updated to some AI/ML model with performance degradation, then DUT can fallback to non-AI/ML mode through model monitoring.**Proposal 1**: RAN4 can discuss tests for model monitoring when LCM framework becomes clear. |
| [**R4-2405706**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405706.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 testting goal 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 testing goal 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 configuration and vary the conditions under the 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 retraining could be beneficial to model delivery/transfer to reduce overhead/latency****Observation 9: Having a different AI/ML model for each different Scenario/configuration and additional condition could increase the UE complexity and storage requirements as well as the overhead of delivery/transfers and the associated overhead/latency.****Observation 10: Options 1-a and 1-b for post deployment testing would impose a large burden for testing, signal overhead and complexity** **Proposal 1**: **RAN4 should study the specification of reference AI/ML models for defining performance requirements for 1 and 2-sided models** **Proposal 2:** **RAN4 should follow the same process of specifying parameters for designing the reference one-sided and two-sided AI/ML models as used in Option 3 for specifying the test decoder. Additionally, it should define a training dataset to ensure consistency and alignment in performance results.** **Proposal 3: Deprioritize Option 1 and focus on Option 2 for performance requirements as a testing goal** **Proposal 4: 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, e.g., channel modes defined for different scenarios (CDL, AWGN, etc.) in TR38.901, Doppler conditions, SNR levels etc.****- Configurations, e.g., number of set A/B beams, input types (wide SSB vs narrow CSI-RS beams), output types (classifier vs L1-RSRP prediction network), performance metrics (for example Option 1 or Option 2 KPI metrics for BM use case), carrier frequency, etc** **Proposal 5: If non-static scenarios/configurations are supported for certain use cases, they can be included as part of generalization tests.****Proposal 6:** **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****Proposal 7: 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 8: To improve the generalization behavior of the model, training with a diverse dataset should be investigated.** **Proposal 9: 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 10: Consider utilizing post-deployment procedures to augment conformance testing for effectively managing performance across all possible deployment conditions/scenarios (which are not tested)**Proposal 11: 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 12: 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 13: 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 14: 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 15: 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.****Proposal 16: 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 17: After fine-tuning UE can update its stored AI/ML models with the new model ID, where this ID can be associated with the training data (which implicitly have the additional conditions) used to fine tune the model.** Proposal 18: 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 conditionsProposal 19: RAN4 should clarify/agree that the side conditions of the testing procedures should remain the same for legacy and AI/ML methods.**Proposal 20: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case** **Proposal 21: RAN4 shall consider the latency requirements for model monitoring input data as well as the establishment of tolerance margin requirements for the specified KPIs for model monitoring per use case****Proposal 22: FFS on how to perform cell level BM performance monitoring when the AI/ML model resides at NW****Proposal 23: For post-deployment validation, 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 training data.** **Proposal 24: For post-deployment validation, 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 25: For post-deployment validation, 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 26: UE updates its capability report with a new model resulted from either: finetuning, model transfer from NW or monitoring to check the consistency of additional conditions. Subsequently UE assigns an ID to the new model that supports the NW additional conditions and shares this ID to NW.** **Proposal 27: 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 28: RAN4 must conduct an analysis for each use case to determine the reliability of using synthetic channels for test data in evaluating models trained on real data.** |
| [**R4-2405737**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405737.zip) | Nokia | **On post-deployment handling:****Observation 1**: The challenge of AI/ML model/functionality flexibility is how to ensure that the device that has passed conformance testing with one version of AI/ML model/functionality can also pass the same test after the upgrade/change of the model/functionality.**Observation 2**: Frequent updates to the models/functionality will prolong the turnaround time of conformance of a model before its deployment in case of conformance testing before any deployment (Option 1).**Proposal 1: Conformance testing for updated AI model/functionality before deployment (Option 1) is feasible in the case of declared centralized changes/updates to the models/functionalities (e.g., based on offline (re)training).****Proposal 2: RAN4 to ensure (and to clarify in Option2: performance monitoring-based post-deployment handling) that in the case of identified performance issues, the fallback/change of the AI/ML model/functionality shall be allowed only to the model/functionality/legacy feature that has passed conformance testing.****Observation 3**: It is hard to conclude just based on monitoring mechanism whether the issue is with the changed model/functionality or due to the challenging conditions.**Proposal 3: RAN4 to consider Option 3 (a new option) for proactive post-deployment handling of AI/ML model/functionality updates: RAN4 to test the procedure when updated/new model/functionality stays inactive in the device before is has passed assessment/verification and can substitute currently active model/functionality.****On data collection requirements:****Proposal 4: In RAN4, requirements on data collection for training, data collection for inference, and monitoring data collection shall be discussed separately.****Proposal 5: RAN4 does not need to discuss accuracy and latency requirement on data collection for training unless training procedures are specified in 3GPP.****Proposal 6: RAN4 to consider monitoring data collection requirements in a use-case specific manner and based on RAN1 design of the corresponding mechanisms.****Observation 4**: For ML-enabled functionalities, even if the internal processing chain in the UE or NW is likely to be more complex and extended with pre/post processing steps before and after the actual ML model, all these steps remain implementation specific and not in the scope of 3GPP specifications.**Observation 5**: The latency of data collection for inference is hard/impossible to verify in RAN4 especially when it includes the stages or interfaces internal to the device and not defined in 3GPP.**Observation 6**: It is more feasible to control the latency in between the radio signal reception by the UE and output/reporting through the standardized 3GPP interfaces, i.e., like in legacy RAN4 core requirements.**Proposal 7:** **RAN4 to focus on the (inference) latency core requirements (i.e., in between the measurements/signalling and reporting) instead of ‘data collection for inference’ and continue the related discussions for each use case separately.****On other general aspects:****Observation 7**: The definitions of AI/ML Model testing and validation introduced in TR 38.843 are not aligned with RAN4. In particular, RAN4 AI/ML-based feature testing cannot be the subprocess of training.**Proposal 8: Add a note in the term definitions (Clause 3.1 of TS 38.843) of AI/ML model testing and AI/ML model validation that they are not applicable in RAN4 context.****Proposal 9: RAN4 to agree and clarify in the TS that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 [1] is applicable only for the testing UE-sided model -enabled use cases.****UL air-interface (conducted/ OTA):**1. Signaling DUT measurements reports
2. Feedback signalling for ML Functionality control messages from DUT

**DL air-interface (conductive/ OTA):**1. Radio access signals
2. Signaling for ML Functionality control messages from TE

Signal generatorinferenceLCMTest configuration/controller AI/ML functionsLCMVerificationProposal 10: RAN4 to agree that the description of the reference block diagram in Figure 7.3.2.3-1 in TR38.843 requires clarifications for at least the following items:* The role and meaning of the ‘LCM’ blocks in the TE and DUT
* The role and meaning of the ‘AI/ML functions’ block in the TE
* The role and meaning of the ‘Inference’ block in the DUT
* The role and meaning of the ‘Verification’ block in the TE
* The signaling/messages assumed on the physical links between the TE/gNB and DUT/UE – see Figure 2.
 |
| [**R4-2405893**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405893.zip) | Keysight Technologies UK Ltd | ***Proposal 1.1:*** *In case LCM is signalling based, RAN4 to define scenarios for inference and generalization in a per-use-case basis not only based in channel model but also include physical parameters that might affect the performance of each AI/ML model.****Proposal 1.2:*** *In case LCM is autonomously carried out by UE, RAN4 to stop discussing about generalization and start to define scenarios for requirements in a per-use-case basis not only based in channel model but also include physical parameters that might affect the performance of each AI/ML model.****Proposal 2****: AI/ML requirements should implicitly/explicitly consider the impact of latency in switching between AI/ML models.****Proposal 3****: Mobility aspects must be considered when defining the CSI prediction scenario.****Proposal 4:*** *In CSI prediction, RAN4 to define requirements for latency related to switching AI/ML models.****Proposal 5****: CSI prediction requirements should include performance of the transition between LOS/NLOS conditions.* ***Proposal 6:*** *CSI prediction requirement to define performance for a range of CSI periodicity values.****Proposal 7****: For post-deployment testing, RAN4 to consider a new option 3: combination of options 1 and 2.****Proposal 8****: Define only data collection latency requirements once it is identified a need to share data between different entities.****Proposal 9:*** *RAN4 to evaluate the need of L1 RSRP data collection accuracy for beam management use case if L1 RSRP is finally the KPI chosen.****Proposal 10:*** *RAN4 to trigger discussions on scenarios to be considered for each use case, including channel model (TDL vs CDL channel model type, channel model parameters, etc.) and physical layer parameters impacting the AI/ML algorithms.*  |

## Open issues summary

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

1. Post deployment testing options
2. Post deployment testing feasibility
3. Testing model/framework
4. CSI testing framework
5. Performance with transitions
6. Inference latency requirements
7. RRC State to consider
8. Testing goals update
9. TR updates : changes to Figure 7.3.2.3-1
10. RRM Impact

### Sub-topic 1-1

*Post-deployment testing options*

Few options for post-deployment testing were agreed in RAN4#110 (R4-24003712). The agreement is reproduced below for convenience.

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

**Agreement:**

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

Some companies proposed to add further options or make some changes. All the options below should be discussed to see if anything is agreeable or needs further changes/clarificaitons

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

* Proposals
	+ Option 1: Add Option 3 (R4-240495 – vivo)

**Option 3: Define the test to verify the performance validation together with model transfer/update**

* + - **validation scheme can be designed by RAN1**
	+ Option 2: Add Option 4 (R4-2405737 – Nokia)

**Option 4: RAN4 to test the procedure when updated/new model/functionality stays inactive in the device before is has passed assessment/verification and can substitute currently active model/functionality.**

* + Option 3: include some mandatory fallback (e.g. to a “baseline” model or older model already tested/validate) as an option
	+ Option 4: As a further option relating to post deployment testing, consider the possibility of capturing model input during testing for later testing of new models: Capture model input during conformance testing for later testing of new models. (FFS the captured data needs to be held completely by the UE vendor) (Ericsson – R4-2405610)
	+ Option 5: others
* Recommended WF

All options need to be discussed

### Sub-topic 1-2

*Post deployment testing feasbility*

As stated in Sub-topic 1-1, several options are listed for further study for post deployment testing. Some companies argue that some options are not feasible, hence, should be dropped.

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

* Proposals
	+ Option 1: Do not pursue Option 1 further, focus on Option 3
	+ Option 2: Adopt the following framework (R4-2404931, Intel)

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

* + - **At least some default AI/ML model (feature or functionality) needs to pass conformance testing and be present in the device.**
		- **Any changes or updates to the ML-enabled functionality or feature, which come from UE-side OTT server, shall be tested by the device vendor against RAN4 requirements before the deployment to the UE is performed. Other model updates may be tested by the device vendor against RAN4 requirements.**
		- **The information on whether AI/ML model update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**
	+ Option 3: Others
	+ Option 4: No need for post-deployment testing, hence, no need to further discuss this issue
* Recommended WF
	+ To be discussed

### Sub-topic 1-3

*Testing environment/framework*

Some companies brought up the issue whether the testing should use static scenarios/environments or dynamic ones.

**Issue 1-3: Testing environment/framework**

* Proposals
	+ Option 1: Use static scenarios as baseline (channel model type and SNR settings are fixed and do not change over the test, specific channel realizations may be dynamic)
	+ Option 2: Non-static scenarios/configuration can be further considered in application to CSI and beam management temporal prediction use cases. The details of models are FFS and may include non-stationary SNR and other conditions.
	+ Option 3: Other options
* Recommended WF
	+ To be discussed

### Sub-topic 1-4

*CSI Testing framework*

Some companies have raised the issue of reusing the CSI legacy framework with relative throughput.

**Issue 1-4: CSI Testing Framework**

* Proposals
	+ Option 1: For CSI compression and CSI prediction, existing RAN4 framework of relative throughput should be reused (PMI reporting
		- legacy performance should be taken as baseline
	+ Option 2: Other options
* Recommended WF

Option 1

### Sub-topic 1-5

*Performance with transitions*

The issue of defining requirements when the environment was changing was brought up

**Issue 1-5: Performance with transitions**

* Proposals
	+ Option 1: RAN4 should define requirements with changing environments (e.g. propagation changing between LOS/NLOS)
	+ Option 2: This discussion should be done as part of generalization on a case by case basis
	+ Option 3: RAN4 should define latency requirements for transitions
	+ Option 4: Other prposals
* Recommended WF

To be discussed

### Sub-topic 1-6

*Inference latency requirements*

**Issue 1-6: Inference latency requirements**

* Proposals
	+ Option 1: RAN4 to focus on the (inference) latency core requirements (i.e., in between the measurements/signalling and reporting) instead of ‘data collection for inference’ and continue the related discussions for each use case separately.
	+ Option 2: No need to discuss inference latency, it will be automatically included in the functionality scheme(e.g. reporting latency/accuracy, performance requirement such as throughput, etc)
	+ Option 3: Postpone this discussion until more details for the procedures defined by RAN1/2 become clear
	+ Option 4: other proposals
* Recommended WF

To be discussed

### Sub-topic 1-7

*RRC state to consider*

**Issue 1-7: RRC state to consider**

* Proposals
	+ Option 1: At least for monitoring data collection, RAN4 will assume RRC\_CONNECTED state for both data generation and reporting.
	+ Option 2: wait for agreements in RAN1/2
* Recommended WF

To be discussed

### Sub-topic 1-8

Testing goals update

The testing goals were already captured in the TR in RAN4#110, some updates are proposed as follows (R4-2405651- Samsung)

* Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.
	+ ~~FFS~~ how to define the specific AI/ML model is provided as use-case specific manner (e.g., a model captured in RAN4 spec as baseline, or a reference model structure agreed for performance alignment)
	+ ~~FFS~~ how to define that the model is properly conducted is provided as use-case specific manner (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)
* Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.
	+ ~~FFS~~ how to define a static scenario/configuration is provided as use-case specific manner (e.g., by defining a related testing dataset based on channel models in TR 38.901)
		- The scenario/configuration shall be determined by “the specific configuration/conditions” associated with the relevant “UE capability of an AI/ML-enabled Feature/FG” under testing
		- The static scenario/configuration shall be maintained unchanged during the test
	+ ~~FFS whether and how to define non-static specific scenarios/configurations~~

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

* Proposals
	+ Option 1: Agree changes to Option 1
	+ Option 2: Agree changes to Option 2
	+ Option 3: Agree both changes
	+ Option 4: make further changes
* Recommended WF

To be discussed . Please provide concrete proposals if Option 4 is prefered

### Sub-topic 1-9

Testing diagram update/clarification

An update to the testing diagram is proposed in R4-2405737(Nokia). It should be discussed whether this is agreeable or not and whether any other changes are needed.

Updated Figure 7.3.2.3-1

**DUT (UE)**

**TE (gNB)**

**DL air-interface (conductive/ OTA):**

1. Radio access signals
2. Signaling for ML Functionality control messages from TE

Figure 2: Reference block diagram for testing ML functionalities enabled with UE-sided models only.
Update for Figure 7.3.2.3-1 in TR38.843 Section 7.3.2.3.

**UL air-interface (conducted/ OTA):**

1. Signaling DUT measurements reports
2. Feedback signalling for ML Functionality control messages from DUT

Signal generator

inference

LCM

Test configuration/controller

AI/ML functions

LCM

Verification

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

* Proposals
	+ Option 1:
		- Proposal 13: RAN4 to agree and clarify in the TS that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 [1] is applicable only for the testing UE-sided model -enabled use cases.
		- Proposal 14: RAN4 to agree that the description of the reference block diagram in Figure 7.3.2.3-1 in TR38.843 requires clarifications for at least the following items:
			* The role and meaning of the ‘LCM’ blocks in the TE and DUT
			* The role and meaning of the ‘AI/ML functions’ block in the TE
			* The role and meaning of the ‘Inference’ block in the DUT
			* The role and meaning of the ‘Verification’ block in the TE
			* The signaling/messages assumed on the physical links between the TE/gNB and DUT/UE
	+ Option 2: Other updates
	+ Option 3: no update needed
* Recommended WF

To be discussed

### Sub-topic 1-10

RRM impact

A table summarizing the possible impact to RRM specifications is presented in R4-2404478. The tables are included below for reference:

**Table 1 General scope of impact on RRM core requirements brought by AI/ML-based beam management**

|  |  |  |
| --- | --- | --- |
| **RRM Req. Category** | **RRM procedure** | **RAN4 impact** |
| IDLE state mobility – Cell selection | Cell selection | No impact |
| IDLE/INACTIVE state mobility – Cell re-selection | Measurement and evaluation of serving cell | No impact |
| Measurements of intra-frequency NR cells |
| Measurements of inter-frequency NR cells/inter-RAT |
| RRC\_CONNECTED state mobility - Handover | NR Handover | No impact |
| Handover to other RATs |
| NR DAPS Handover |
| NR Conditional Handover |
| NR Handover with PSCell |
| RRC\_CONNECTED state mobility - RRC Connection Mobility Control | SA: RRC Re-establishment | No impact |
| Random access |
| SA: RRC Connection Release with Redirection |
| Timing | UE Transmit timing | No impact |
| UE Timer accuracy |
| Timing advance |
| Cell phase synchronization accuracy |
| Maximum transmission timing difference |
| Maximum receive timing difference |
| Signaling characteristics  | SSB /CSI-RS based Radio Link Monitoring | Discussion and Decide |
| Interruption | No impact |
| Scell Activation/Deactivation Delay |
| UE UL carrier RRC reconfiguration delay |
| Link recovery procedures (SSB/CSI-RS based beam failure indication, L1 indication, Candidate beam detection) | Discussion and Decide |
| Active BWP switch delay | No impact |
| Active TCI state switching delay | Discussion and Decide |
| Uplink spatial relation switch delay |
| UE-specific CBW change | No impact |
| Pathloss reference signal switching delay | Discussion and Decide |
| Active downlink TCI switching delay for unified TCI |
| Active uplink TCI switching delay for unified TCI |
| Functionality/Model monitoring procedure | FFS new requirements |
| Functionality/Model selection/activation/deactivation/fallback delay |
| Active model switch delay |
| Measurement procedure | NR intra-frequency measurements (Number of cells and number of SSB, measurement reporting requirements, etc) | Discussion and Decide |
| NR inter-frequency measurements | No impact |
| Inter-RAT measurements |
| L1-RSRP measurements for reporting | Discussion and Decide |
| Cross link interference measurements | No impact |
| L1-SINR measurements for reporting |
| Measurement performance requirements – NR measurements | Intra-frequency RSRP accuracy requirements for FR1/FR2 | Discussion and Decide |
| Inter-frequency RSRP accuracy requirements for FR1/FR2 | No impact |
| RSRP measurement report mapping | Discussion and Decide |
| Intra-frequency RSRQ accuracy requirements for FR1/FR2 | No impact |
| Inter-frequency RSRQ accuracy requirements for FR1/FR2 |
| RSRQ report mapping |
| Intra-frequency SINR accuracy requirements for FR1/FR2 |
| Inter-frequency SINR accuracy requirements for FR1/FR2 |
| SINR report mapping |
| L1-RSRP accuracy requirements for FR1/FR2 | Discussion and Decide |
| L1-SINR accuracy requirements for FR1/FR2 | No impact |

**Table 2 General scope of impact on RRM core requirements brought by AI/ML-based positioning accuracy enhancement**

|  |  |  |
| --- | --- | --- |
| **RRM Req. Category** | **RRM procedure** | **RAN4 impact** |
| Signaling characteristics | Functionality/Model monitoring procedure | FFS new requirements |
| Functionality/Model selection/activation/deactivation/fallback delay |
| Active model switch delay |
| Measurement procedure – NR measurements for positioning | RSTD measurements | Discussion and Decide |
| PRS-RSRP measurements |
| UE Rx-Tx time difference measurements |
| E-CID measurements |
| PRS-RSRPP measurements |
| Measurement performance requirements – NR measurements | RSTD measurements | Discussion and Decide |
| PRS-RSRP measurements |
| UE Rx-Tx time difference measurements |
| PRS-RSRPP measurements |
| Measurements performance requirements for NR gNB | UL-RTOA | Discussion and Decide |
| gNB Rx-Tx time difference |
| UL SRS RSRP measurement |
| Timing advance |
| UL SRS RSRPP measurement |

**Issue 1-10: RRM Impact**

* Proposals
	+ Option 1: Please provide comments on whether this is expected impact or other requirements might also be impacted
	+ Option 2: RAN4 starts the LCM core requirements discussion focusing on the procedures from the common set which are relevant for both functionality- and model-based LCM, so that whichever option is selected in the end, this initial work will still be relevant.
	+ Option 3: The initial RAN4 discussion on the LCM core requirements can include: the requirements scope, type of the requirements, and parameters in the requirements, with the understanding that the actual values in the requirements may differ for the functionality-based and model-based LCM.
	+ Option 4: RAN4 to discuss and define upon the need requirements for switching between the AI mode and the non-AI mode.
	+ Option 5: Others
* Recommended WF

To be discussed

**Issue 1-11: RRC state to consider**

* Proposals
	+ Option 1: At least for monitoring data collection, RAN4 will assume RRC\_CONNECTED state for both data generation and reporting.

# 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** |
| [**R4-2404143**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404143.zip) | Korea Testing Laboratory | **Observation 1: Upon the deployment circumstances, the number of TX beams can vary such as 8 to 64 downlink Tx beams (max number of available beams) at the NW side. Other values, e.g., 256 not precluded.****Observation 2: The number of UE beams can vary by 4 or 8 downlink Rx beams (the max number of available beams) per UE panel at UE side. Other values, e.g., 16 not precluded.** **Observation 3: It is essential to provide the spatial correlation for the test setup environment for BM-Case1.** **Observation 4: The testability and interoperability for BM-Case1 and 2 shall have sustainability for the various deployment circumstances of the gNB antenna configuration.*****Proposal 1: Study and specify the single testing environment to keep sustainability for the various deployment scenarios.******Proposal 2: The following options should be considered**** ***Option 1: Study whether or not the existing FR2 OTA chamber has a sustainable testing environment providing the spatial property for BM-Case1 and 2.***
* ***Option 2: Study and specify the minimum required number of TX beams and the test setup environment to evaluate the inference algorithm such as 4x4 or 2X4 or 1X8 TX Beams where other configuration is not precluded for the test setup that shall have single and independent evaluation environment.***

**Observation 5: In the evaluation, UE rotation is modelled for BM-Case2 with a rotation speed in all three rotational axes, with the rotational direction chosen uniformly at random among the three axes.*****Proposal 3: The following issues should be considered**** ***Study on the testing BM-Case2 whether or not the inference algorithm is independent of the rotational direction of UE.***
	+ ***Study the test setup on the existing FR2 OTA setup as baseline.***
	+ ***Otherwise, study a test environment considering the rotational direction of UE***
 |
| [**R4-2404152**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404152.zip) | MediaTek inc. | **Observation 1**: Considering the UE-sided model and FR2 test, the data for inference are obtained by measuring the beams in set B transmitted from TE. Therefore, the measured L1-RSRP will anyway contains measurement errors.**Observation 2**: Either ideal L1-RSPR or measured L1-RSRP could be used for offline training.**Observation 3**: It will be easy for training data collection if we use the ideal L1-RSRP generating from channel emulator.**Observation 4**: During the test or in the field, the inference input to the AI/ML model is measured L1-RSRP, not ideal L1-RSRP. If the AI/ML model is trained with ideal L1-RSRP, it may cause performance degradation due to the mismatch between the data for training and inference.**Observation 5**: Either ideal L1-RSPR or measured L1-RSRP could be used for performance evaluation.**Observation 6**: Even though the measured L1-RSRP in set A at the TE side for performance evaluation can be obtained by the report from the UE side, it will complicate the whole testing process as UE needs to measure all beams in set A.**Observation 7**: From our simulation results, there is a large inference performance difference between ideal L1-RSPR and measured L1-RSRP.**Proposal 1**: Based on observations 1~7, we suggest RAN4 to investigate the testability of ideal L1-RSRP and measured L1-RSRP in the BM performance test.**Observation 8**: From our simulation results, there is a non-negligible difference between different measurement error modelling when considering measured L1-RSRP.**Proposal 2**: If RAN4 agreed to use measured L1-RSRP in BM test, RAN4 should conclude the error modelling when deriving the requirements.**Observation 9**: Different TE and UE implementation may impact the inference performance of AI/ML BM.**Proposal 3**: RAN4 to discuss how to verify the assumptions used for simulation to reflect the real correlation and measurement noise for set A and set B when UE is tested in an anechoic chamber.**Observation 10**: Only DL Tx beam prediction is in scope and Rx beam prediction is not for both spatial-domain and temporal beam prediction. **Proposal 4**: The TCI state QCL to an RS that is not in Set B is unknown if no L1 measurement is performed by UE within 1280ms before TCI state activation. **Proposal 5**: The TCI state which is QCL to an RS in Set B is known:* if the corresponding predicted beam is reported in 1280ms before the TCI state switch command and SNR of the RS is above -3dB for spatial-domain beam prediction.
* if the last observation occasion is within 1280ms before the TCI state switch command and SNR of the RS is above -3dB for temporal beam prediction.
 |
| [**R4-2404215**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404215.zip) | Qualcomm Incorporated | **Observation 1: An effective test case for beam prediction use case BM-Case1 requires the following:*** **Sufficient randomness and variation in time and spatial domain of L1-RSRP has to be emulated in the test**
* **Support emulation of DL Tx beam sweeping with enough number of Tx beams in Set B and Set A**

**Observation 2: RAN4 L1-RSRP and other measurement test configurations support only:*** **Deterministic and static power configuration on each AoA**
* **Emulation of received signals from up to 2 AoAs**

**MIMO OTA testing environment defined in Rel-17 support only:*** **Emulation of CDL channel with fixed DL beam to the channel**
* **Emulation of received signals from up to 6 AoAs/probes, and the coverage (in terms of area on the sphere) by the 6 probes is very limited and sparse based on TR 38.827 Table 6.2.3-1, most of them are in Theta [0,30] and Phi [0,120], only one “mirror” point at negative Phi region.**

**Observation 3: The following conditions contribute to randomness and variation in time and spatial domain of L1-RSRP*** **Propagation conditions as a function of (1) AoD of the Tx beam (2) AoA of the Rx beam (3) fading condition, e.g., a CDL channel**
* **Tx beamforming gain on the AoDs in the propagation conditions**
* **UE movement**

**Proposal 1: DL Tx beam sweeping for BM-Case 1 test can be emulated by the time-varying input power to the CDL channel model emulated on top of MIMO OTA test environment.****Proposal 2: Received power emulation can follow the formula below:**$$P\_{r}\left(AOA\_{k},i\_{Tx}\right)=\sum\_{AOA\_{j}=(or≅)AOA\_{k}}^{}P\_{t}×G\_{Tx}\left(i\_{Tx},AOD\_{j}\right)×G\_{path,j}\left(AOD\_{j},AOA\_{j}\right)$$**Where** $G\_{Tx}\left(i\_{Tx},AOD\_{j}\right)$ **is Tx beamforming gain as a function of beam index and AOD, and channel gain (from CDL channel model) on each path,** $G\_{path,j}\left(AOD\_{j},AOA\_{j}\right)$**, is a function of AOD and AOA of the path.****Proposal 3: Tx beam sweep can be emulated based on the following formula by the probes on AoAk:**$$P\_{r}\left(AOA\_{k},t\right)=\sum\_{AOA\_{j}=(or≅)AOA\_{k}}^{}P\_{t}(t)×G\_{Tx}\left(i\_{Tx}(t),AOD\_{j}\right)×G\_{path,j}\left(AOD\_{j},AOA\_{j},t\right)$$**Where AoAk is the AoA of probe *k*, channel gain at time *t* denoted by** $G\_{path,j}\left(AOD\_{j},AOA\_{j},t\right)$ **sampled from CDL model, and Tx beam used in time *t* denoted by** $i\_{Tx}\left(t\right)$ **based on the beam sweeping RS transmission pattern. Note that the probe power configuration can be verified by comparing the probe power as a function of time and the distribution of** $P\_{r}\left(AOA\_{k},t\right)$**.****Observation 4: UE received power when DL Tx beam** $i\_{Tx}$ **is transmitting can be formulated as:**$$P\_{r,UE}\left(t\right)=\sum\_{\left\{AOA\_{k}\right\}}^{}\left(G\_{Rx}\left(i\_{Rx},AOA\_{k}\right)×\sum\_{AOA\_{j}=(or≅)AOA\_{k}}^{}P\_{t}×G\_{Tx}\left(i\_{Tx}(t),AOD\_{j}\right)×G\_{path,j}\left(AOD\_{j},AOA\_{j},t\right)\right)$$**However, this quantity is unknown to TE since TE doesn’t have access to the** $G\_{Rx}\left(i\_{Rx},AOA\_{k}\right)$**.****Observation 5: One candidate resolution to ground truth availability is explained in the following. We can configure the test and TE channel emulator (fader) so that the channel from t0 to t1 is the same as channel from t1 to t2**$$G\_{path,j}\left(AOD\_{j},AOA\_{j},t\right)|\_{t=t0}^{t=t1}= G\_{path,j}\left(AOD\_{j},AOA\_{j},t\right)|\_{t=t1}^{t=t2}$$**And so does the beam sweeping** $$G\_{Tx}\left(i\_{Tx}(t),AOD\_{j}\right)|\_{t=t0}^{t=t1}=G\_{Tx}\left(i\_{Tx}(t),AOD\_{j}\right)|\_{t=t1}^{t=t2} $$**Then we can consider “measured”** $P\_{r,UE}\left(t\right)$ **from t1 to t2 as ground truth, and compare it against predicted** $P\_{r,UE}\left(t\right)$ **from t0 to t1 by designing the test as t0 to t1, only set B is transmitted; and in t = t1 to t2, both set A and B are transmitted and TE request measurement report (for both set A and B).****Proposal 4: The following issues should be considered when defining the beam prediction accuracy requirements*** **Consistency between training and testing data (from the perspectives of beams in Set B and Set A, physical characteristics of gNB antenna etc.) should be guaranteed by signaling conveyed to UE.**
* **The impact of size and composition of Set B and Set A on accuracy requirement.**

**Proposal 5: RAN4 can consider UE rotation by leveraging the physical motion control of the DUT holder. With UE rotation, the probe coverage becomes** $\bigcup\_{t in test time}^{}\left\{AOA\_{k}(t)\right\}$**, where AoAs are time varying functions based on UE rotation (note that AoA is represented in UE coordination system). The probe functions for emulating Tx sweeping remain the same as proposal 3 except that AoAs become functions of time:** |
| [**R4-2404427**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404427.zip) | CATT | **Observation 1: Beam prediction accuracy is more appropriate to be metrics/KPIs for BM use cases, while the predicting RSRP optionally depends on the capability of the deployed AI/ML models.** **Proposal 1: RAN4 wait for RAN1 conclusion on report contents for inference and comeback to discuss the details of metrics/KPIs, if any.** **Proposal 2: RAN4 discuss whether to tight accuracy requirements for AI/ML enabled BM in performance part based on evaluation results.** **Proposal 3: Ground truth is the Top-K reference/target beam ID(s) that are known to TE, if beam prediction accuracy is chosen as metrics (Observation 1).** **Proposal 4: Dataset can be built based on the channel models in TR 38.901 and a data format can also be standardized. Dataset can be updated by adding data, e.g., field data or data obtained by other methods, which have the standardized format from volunteers.** |
| [**R4-2404480**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404480.zip) | CAICT | **Proposal 1: Suggest to discuss NW-side model and UE-side model separately.** |
| [**R4-2404484**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404484.zip) | LG Electronics Inc. | ***Proposal 1***: Support option 4. Combination of option 1 and option 3 in TR are our preference.***Proposal 2***: Support option 3. Accuracy requirements cannot be tightened. If necessary, it can be measure multiple time to get more accurate results. ***Proposal 3***: Support option 2 and 3. Only use AWGN channels. It looks like that there are no difference between option 2 and option 3. We think option 2 and option 3 can be combined. |
| [**R4-2404573**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404573.zip) | Xiaomi | **Observation 1: In legacy non-AI based L1-RSRP accuracy requirement, both absolute and relative accuracy requirements are defined.****Observation 2: For non-AI based absolute L1-RSRP accuracy test metric, ideal RSRP is not a value but a range considering minimum and maximum RX beamforming gain. Ideal RSRP is determined by transmit signal power at TE side and RX beamforming gain range of UE.****Observation 3: For non-AI based absolute L1-RSRP accuracy test metric, RSRP difference is measured RSRP with ideal RSRP for the same beam index. No best beam selection functionality is tested.****Observation 4: Relative L1-RSRP accuracy is designed to verify baseband performance which can reduce or get rid of the impact of large RX beamforming gain range. Relative accuracy is the measured L1-RSRP difference between two RSs.****Observation 5: For option 1, RSRP prediction accuracy focus about the RSRP difference for the same beam.****Observation 6: Both option 2 and 3 can be classified into beam prediction accuracy. Option 2 refers to beam index difference while option 3 refers to RSRP difference between predicted best beam and ideal best beam.** **Proposal 1: Option 3 needs to be updated by considering both upper bound and lower bound.*** **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 and smaller than the RSRP of the strongest beam + x dB,**
	+ **Related measurement accuracy can be considered to determine x**

**Observation 7: RSRP prediction accuracy requirement can verify both RSRP accuracy and beam prediction accuracy only under certain conditions.****Proposal 2: RAN4 needs to at least define beam prediction accuracy requirement and discuss which test metric to be chosen, e.g. based on beam index difference or RSRP difference, or both.****Proposal 3:** **AWGN channel is not suitable for beam prediction test since there is no any uncertainty in spatial domain. Random channel should be chosen to verify prediction functionality.****Observation 8: For Option 1, it’s possible that TE knows ideal L1-RSRP at least for BM case-1 under some conditions.****Observation 9: For option 2, TE may not know the ideal best TX beam index for some cases.****Observation 10: For option 3, it’s more challenging for TE to know the corresponding L1-RSRP for the strongest TX beam.****Proposal 4: Regarding to whether ideal value can be known or not, we propose:*** **RAN4 to discuss for each test metric respectively**
* **RAN4 to discuss for BM case-1 and BM case-2 respectively**
* **RAN4 to discuss possible condition if ideal value can be known**

**Observation 11: If measured value is used as ground truth, it’s the requirement from another aspect. It’s some kind of relative accuracy compared with measurement. There are both pros and cons.****Proposal 5: RAN4 to discuss whether measured value can be used as ground truth.** |
| [**R4-2404721**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404721.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.*** * ***The motivation is that for Top-K/1 (%), with larger value of K, it is more robust to the measurement error. It is suggested to check whether there is K with reasonable value, e.g. K between 1 to 5, which is roubust enough and the performance degredation due to measurement error is minimum.***

***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.******Proposal 3: for beam management, it is proposed to take RSRP accuracy and beam prediction accuracy as metrics for requirements/tests.*** |
| [**R4-2404946**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404946.zip) | vivo | **Observation 1: Both predicted Top-K beam and predicted RSRP are supported to report in RAN1.****Proposal 1: RAN4 to confirm to use the following prediction accuracy as the Metrics/KPIs for Beam prediction requirements/tests:*** **Top-K/1 (%): the percentage of "the Top-1 strongest beam is one of the Top-K predicted beams**

**Proposal 2: For beam prediction requirements/tests, RSRP accuracy can be the additional Metrics/KPIs when predicted RSRP is required to report. Specifically, RSRP accuracy pertains to the scenario where the top predicted AI/ML Beam ID is the same with legacy’s/genie beam****Proposal 3: RAN4 to further discuss the impact of measurement accuracy on prediction accuracy after the metrics/KPIs for Beam prediction accuracy requirements/tests are determined****Proposal 4: For generalization on beam management, RAN4 to firstly figure out the key factors (e.g., the beam number set A/set B, UE mobility and the codebook configuration of base station) that affect performance and further discuss how many representative scenarios to be tested in the test coverage****Proposal 5: RAN4 to further study the testing feasibility and possible methodology for AI/ML beam management in Rel-19 study item on NR FR2 OTA testing enhancement** |
| [**R4-2405021**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405021.zip) | Ericsson | **Observation 1: Among options in test metrics, the RSRP accuracy cannot be identified as a test metric without beam information.****Observation 2: The approximated predicted RSRP values among different beams might cause a misjudgment and impact the beam prediction accuracy.****Observation 3: A tolerance margin to avoid the beam prediction ambiguity due to RSRP proximity is needed to be defined as a supplementary information for option 2. As reference, ‘beam prediction accuracy (%) with 1dB margin for Top-1 beam’ for evaluating the performance of AI/ML in beam management is defined in RAN1.****Observation 4: The conformance test shall avoid uncertainty/variance of the relationship between beam information and RSRP, which can reduce the test complexity including post processing and test case number, RAN4 shall identify the issue in test configuration and definition after the test metric is identified.****Observation 5: If the UE makes decisions on model selection/activation/deactivation/switching/fallback, then there may be a need for requirements to ensure performance continuity, e.g., specify the needed interruption time.****Observation 6: If the network indicates to the UE to do LCM operations, there may be a need for requirements such as activation/deactivation time, interruption time etc, e.g., after receiving a command to operate the model in UE, the UE shall complete the operation in a required time span.****Observation 7: The application of AI/ML model, compared with the legacy non-AI method, may cause worse network performance when the measurement error during training and inference phase is large.** **Observation 8: Side condition of -3dB for L1-RSRP measurement requirement, compared with no limitation, may degrade AI/ML performance.** **Observation 9: To define the generalization, specially from test perspective, we consider the below options:*** **Option 1: The AI/ML model is fixed, for various scenario**
* **Option 2: The AI/ML model is semi-fixed (no re-training is needed), for various scenario**
	+ **Option 2.1: AI/ML model can adapt various scenario without any external signaling**
	+ **Option 2.2: AI/ML model can adapt various scenario with indication by any external signalling, e.g., UE has more information due to changes of its radio conditions, based on which the UE can trigger change/switch model for the scenario.**
* **Option 3: The AI/ML model is changed, i.e., re-trained, for various scenario**

**Proposal 1: For test purposes, the predicted RSRP should be reported together with beam information such as beam ID.****Proposal 2: Option 3 in test metrics should be rephrased, e.g., merged/covered by the final agreements depending on Option 1 and Option 2, or deprioritized.****Proposal 3: From verification in RAN4 perspective, Performance metric(s) Alt.1: Beam prediction accuracy and Alt.4: The L1-RSRP difference shall be prioritized.** **Proposal 4: It’s preferred to align the performance monitoring metrics and that for test metrics.****Proposal 5：For performance monitoring at UE side, there may be a need to discuss requirements on the accuracy of reporting that is sent by the UE to the network. Potentially also measurement duration requirements may need to be considered. Wherein, the report including:** * **performance metric(s) calculated by the UE, either reports it to NW or reports an event to NW based on the performance metric(s), or**
* **measurement report and prediction report**

**Proposal 6: RAN4 shall clarify the understanding of measurement for performance monitoring, it may rely on RAN1 agreements.** * **One fundamental understanding is the measurement in monitoring shall not occupy RS resources as legacy measurements or shall not extend legacy measurements, for saving power consumption.**
* **Any necessary dependence between measurement for performance monitoring and model training.**

**Proposal 7: RAN4 shall determine whether the worse network performance is acceptable when AI/ML model provides worse performance than legacy measurement.****Proposal 8: RAN4 shall determine whether to tighten the range of acceptable RSRP accuracy of AI/ML model.****Proposal 9: RAN 4 is expected to investigate the side condition for AI/ML model, e.g., design different absolute or relative accuracy requirements for various side conditions.****Proposal 10: Latency requirement in AI/ML model shall be pursued in RRM specification, which may comprise below aspects:*** **Inference latency**
* **Monitoring latency**
* **LCM latency**

**Proposal 11: The feasibility of creating synchronized CDL models with multiple beams (Set A/B) with up to 2 beam probes shall be studied.****Proposal 12: Static scenario/configuration may be deprioritized since the capability (robustness) of AI/ML model cannot be exploited with static scenario/configuration.****Proposal 13: RAN4 shall check whether the AI/ML model fulfills generalization, if the AI/ML model is semi-fixed (no tr-training is needed) but can adapt various scenarios based on determination by UE itself without any external signaling, to guarantee performance in various scenario.**   |
| [**R4-2405149**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405149.zip) | Huawei,HiSilicon | ***Proposal 1:*** For NW-sided beam prediction, the requirement of RSRP measurement reporting reuses RAN4 legacy.***Proposal 2:*** For UE-sided beam prediction, take Option 1 as baseline to ensure that the received RSRP is larger than a pre-defined threshold, if TE uses the predicted beam for DL transmission.***Observation 1:*** The test metric in Option 2 is not applicable if the beam measurement accuracy tolerance is not considered.***Proposal 3:*** For UE-sided beam prediction, deprioritize the test metric of beam prediction accuracy in Option 2 |
| [**R4-2405185**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405185.zip) | OPPO | **Proposal 1: Regarding the model input tests, performance requirement on beam measurement accuracy should be considered.****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 is the only available KPI accordingly,**
* **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 should be utilized as RAN4 test KPIs.**

**Proposal 3: UE can monitor the performance of AI/ML based CSI model/functionality through RSRP accuracy or Beam prediction accuracy, depends on the progress in RAN1.****Proposal 4: Stability of the performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes.****Proposal 5: For BM testability, RAN4 need to:**1. **Clarify the limitations regarding the FR2 beams or probes that TE vendors could support**
2. **To determine the test setup for BM, a potential approach could be:**
* **Assuming that TE supporting 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 X probes that could be supported by TE vendors**
* **FFS how to map the X AoA and corresponding X pathloss to different beams in BM set A/B**

1. **Consider how to ensure that the BM model constructed on the DUT side can match(or approximate match) and be utilized in the testing environment on the TE side**
 |
| [**R4-2405188**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405188.zip) | ZTE Corporation | **Observation 1: The UE shall report the predicted RSRP corresponding to predicted beams ID, then the TE will check whether the range of predicted RSRP of Top-K beams includes the RSRP value of strongest beam.****Observation 2: The UE will report the predicted beam ID, then the TE will check whether predicted Top-K beams ID includes the strongest beam.****Proposal 1: The RSRP accuracy of Top-K or Top-1 predicted beams shall be used as performance metrics for AI/ML based beam management.****Proposal 2: Top-K or Top-1 beam ID prediction accuracy shall be used as performance metrics for AI/ML based beam management.****Observation 3: The different assumptions may impact the related requirements such as measurement error.****Proposal 3: RAN4 shall consider the different assumptions when specify the performance requirements for AI/ML based beam management**.**Proposal 4: RAN4 shall consider the TE capability since it may bring the high complexity and high cost.****Observation 4: In the legacy test, the AWGN channel is assumed as the propagation condition in the test and the upper and lower bound of ground truth values have been confirmed. However, if the fading channel is considered how to define the upper and lower bound.** |
| [**R4-2405652**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405652.zip) | Samsung | *RAN4 Core Requirements for Supporting NW-sided Model***Observation 1: From RAN1 perspective, the enhancement on UE measurement and reporting (i.e., increased number of reported RSRPs and the increased number of measured beams/CSI-RS resources) could be introduced for Rel-19 AI-BM for data collection.****Observation 2: For data collection for NW-sided model inference, the enhancement on UE beam reporting (i.e., the report of more than 4 beam related information) in L1 signaling could be introduced for Rel-19 AI-BM.****Observation 3: From RAN1 perspective, the approaches for overhead reduction could be introduced for Rel-19 AI-BM.****Observation 4: For data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, L1 signaling or higher-layer signaling could be used to report the collected data.** **Proposal 1: RAN4 shall introduce the necessary core requirement on supporting data collection for NW-side AI/ML model of BM-Case1 and BM-Case2, by considering the new aspects including:** 1. **enhancement on UE measurement and reporting (e.g., beam reporting for more than 4 beam for model inference in L1 signaling);**
2. **overhead reduction on UE reporting;**
3. **L1 or higher-layer signalling design.**

**Proposal 2: For the necessity/feasibility of tightening measurement accuracy requirement for AI-BM NW-sided model:** 1. **RAN4 shall only focus on the feasibility part of tightened measurement accuracy;**
2. **RAN4 discussion shall consider at least (i) absolute/relative RSRP used for AI/ML and (ii) how to achieve tightened measurement accuracy in UE side.**

**Proposal 3: Except UE beam reporting enhancement for model inference, no RAN4 impact is expected for model inference for BM-Case1 and BM-Case2 with a NW-side AI/ML model.** **Proposal 4: No RAN4 impact is expected for performance monitoring for BM-Case1 and BM-Case2 with a NW-side AI/ML model.** **Proposal 5: For LCM of NW-sided model, RAN4 shall further study the necessary RAN4 requirements, depending on RAN2-introduced signalling/mechanism, if any.** *RAN4 Core Requirements for Supporting UE-sided Model***Observation 5: The potential interaction(s) between UE and NW for data collection for UE-sided model could be: (1) UE reporting of the supported/preferred DL RS configuration; (2) trigger/initiating data collection; (3) assistance information from NW to UE for data collection.****Proposal 6: RAN4 shall not define any requirement for data collection for UE-sided model training.****Proposal 7: The consistency/association of Set B beams and Set A beams across training and inference shall be guaranteed in RAN4 core requirement as the NW-side additional assistance information for UE-sided model testing for inference and monitoring.** **Proposal 8: RAN4 requirement for model inference with UE-side AI/ML model needs to be specified, by considering the (potentially) new mechanisms, including (1) indication of the associated Set A from network to UE, (2) beam indication from network for UE reception (if new indication introduced), (3) (3) beam information on predicted Top K beam(s) among a set of beams and other information.****Proposal 9: For metrics for beam management requirements/tests, Option 2 (beam prediction accuracy) shall be considered as baseline.****Observation 6: For UE-sided model, RAN4 requirement on performance monitoring will be specified based on RAN1 conclusion on performance metric(s) and benchmark/reference for the performance comparison.** **Proposal 10: For different types of performance monitoring for UE-sided model, the necessity of RAN4 requirement is provided as:** 1. **Type 1, Option 1 (NW-side performance monitoring): The necessity of RAN4 requirement on data collection for monitoring is not significant, because it is similar to data collection for other purposes.**
2. **Type 1, Option 2 (UE-assisted performance monitoring): RAN4 requirement on data collection for monitoring can be specified to test the accuracy of performance metrics calculated by UE.**
3. **Type 2 (UE-side performance monitoring): No RAN4 requirement is needed because no UE feedback will be performed.**

**Proposal 11: For LCM for UE-sided model, the necessity of RAN4 requirement is provided as:** 1. **LCM performed by the gNB: RAN4 requirement is needed.**
2. **LCM performed by the UE: No RAN4 requirement is needed.**

*Summary of Proposals for AI-BM Core Requirement***Proposal 12: The analysis on the expected RAN4 requirement impact is summarized as:**

|  |  |  |  |
| --- | --- | --- | --- |
| UE/NW | Operations | Expected RAN4 Requirement Impact | Samsung Proposal |
| NW-sided model | Data collection | (1) RAN4 requirement for enhancement on UE measurement and report to support data collection (i.e., the contents of the collected data) | Proposal 1 (Necessary core requirement for data collection)Proposal 2 (FFS the feasibility part of tightened measurement accuracy in RAN4) |
| (2) RAN4 requirement for supporting overhead reduction (the omission/selection of collected data, and/or the compression of collected data), if introduced. |
| Inference | N/A (gNB-Implementation based)  | Proposal 3 (No RAN4 impact expected) |
| Perf. monitoring | N/A (gNB-Implementation based)  | Proposal 4 (No RAN4 impact expected) |
| LCM | (3) RAN4 requirement impact, by considering the design of signaling/mechanism(s) for LCM | Proposal 5 (FFS, depends on RAN2 introduced signaling/mechanism if any)  |
| UE-sided model | Data collection | N/A at least for data collection for model training (UE-Implementation based)  | Proposal 6 (No RAN4 requirement for training) |
| Additional assistance information | (4) Impact on RAN4 requirement to ensure consistency/association between training and inference regarding NW-side additional assistance information for inference at UE.  | Proposal 7 (Consistency/association shall be guaranteed in RAN4 requirement) |
| Inference | (5) RAN4 requirement for supporting model inference. | Proposal 8 (Necessary core requirement for model inference with new RAN1 mechanisms)Proposal 9 (Option 2 (beam prediction accuracy) for metrics for beam management requirement/tests.) |
| Perf. monitoring | (6) RAN4 requirement for supporting performance monitoring. | Proposal 10 (RAN4 requirement on Type 1 (Option 2) performance monitoring is required.) |
| LCM | (7) RAN4 requirement for supporting LCM. | Proposal 11 (RAN4 requirement for LCM performed by the gNB is required) |

*Testability Issues for AI-BM***Proposal 13: Testability study on testing AI-BM shall be based on one of the existing FR2 OTA chamber systems:** **- Option 1: Direct Far Field (DFF) OTA chamber for FR2 RRM testing****- Option 2: 3D Multi-Probe Anechoic Chamber (MPAC) for FR2 MIMO OTA testing****Proposal 14: FFS the feasibility of DFF chamber and 3D MPAC in terms of:** **- The necessity of evaluate RX beam management in OTA chamber;****- The feasibility to generate the required channel model (for certain TX beambook) with limited number of probes.**  |
| [**R4-2405707**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405707.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 testing purposes and defining requirements for BM, a UE can pass a test under one test metric but it could fail the other Options. There are many combinations of pass/fail or fail/pass** **Observation 6: Ensuring identical beam IDs between legacy/genie system and AI/ML models for RSRP comparisons, as well as maintaining RSRP accuracy across multiple beams, poses a challenge for the Option 3 test metric. (that is, it fails to test for those)****Observation 7: For option 2 test metric, using the beam ID prediction accuracy as the KPI for validating the test could lead to testing issues since RSRP accuracy couldn’t be guaranteed or tested (since we only test beam ID)** **Observation 8: RAN4 needs to capture a different test metric or a combination of Options 1,2,3****Observation 9: 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 10:** **For training data based on real measurements, the** **quality of training data depends on RF impairments, and other noise sources. There is tradeoff between training data quality and generalization performance.** **Observation 11:** **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 12:** **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 13: 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 14: Achieving consistency between training and inference by model monitoring could result in delays and increased complexity in model management for BM use case****Observation 15: 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 16: 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.****Proposal 1**: **RAN4 should study the specification of reference AI/ML models, training procedure, and training data for defining performance requirements for BM use case while considering limitations on model complexity. We provide a reference table for initiating the discussion.****Proposal 2:** **RAN4 should investigate the feasibility of the current FR2 OTA-based test procedure to capture random fading in both spatial and temporal domains, similar to CDL, Uma, etc. Additionally, considering the incorporation of UE rotation can help model randomness from the perspective of the Angle of Arrival (AoA)****Proposal 3: If the current FR2 OTA-based test procedure proves to be not feasible, consider using CDL and Uma channels for testing in the BM AI/ML use case****Proposal 4: For testing purposes and defining requirements for BM, RAN4 should clarify that RSRP accuracy pertains to the scenario where the top predicted AI/ML Beam ID is the same with legacy’s/genie****Proposal 5: The KPI test metric should take into account the UE capability for functionality identification. RSRP accuracy should not be a relevant test if the model type ouput of a model is indicated to be Beam ID prediction (classifier based)** **Proposal 6: For the RSRP accuracy test metric, if the predicted beam ID that matches the genie ID doesn’t belong to the top-K best beam IDs, the test should fail irrespectively of the RSRP similarity between predicted and measured.****Proposal 7: There are issues with for Option 3 test metric for beam management requirements/tests. RAN4 to revise option 3 to address the issues discussed or propose a different test metric.****Proposal 8: If a UE is only tested for option 2, the NW should request another beam sweep for obtainining measurements based on the predicted beam IDs. In this way the RSRPs are measured in the second round and the RSRP accuracy between predicted and measured is not an issue any more** **Proposal 9: Part of UE capability signaling should indicate the testing option used for the UE. This information enables the network to determine whether to initiate another sweep of beams for measurements. For instance, if the UE has already been tested for RSRP accuracy, the network can skip additional UE measurements.****Proposal 10: It is important to differentiate predicted beam reporting and measured beam reporting and for that purpose NW could configure separate CSI reporting configurations. Various CSI-Report Configs can be configured for the UE, with one designated for legacy measured beam reporting and another for predicted beam reporting****Proposal 11: For testing purposes and defining requirements for BM, RAN4 to specify a new KPI that simultaneously captures similarity across both RSRP and Beam ID domains as described above. We call this Option 4.****Proposal 12: RAN4 to investigate the source of training data for BM by taking into consideration the advantages and disadvanges of all the considered options** **Proposal 13:**  **For training data based on real measurements, for BM-Case1 and BM-Case2, RAN4 should study the impact of legacy L1-RSRP measurement accuracy requirements (accuracy of training data) as well as the quality of those data on the performance and generalization of AI/ML based BM****Proposal 14:**  **RAN4 to jointly consider the tradeoffs for selecting the source of training and testing data for BM use case.****Proposal 15: 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 16: 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 17: 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 18:** **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.****Proposal 19:  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 20: Investigate the feasibility of enhancing the generalizability of the AI/ML model and reducing the number of AI/ML models and the testing burden for the beam management case by supplementing the core AI/ML input signals with both network (NW) and UE auxiliary information signals integral to its inference engine****Proposal 21: 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.****Proposal 22: 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 23: 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 24: 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 25: 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 26: 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.**  |
| [**R4-2405794**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405794.zip) | Nokia | **Observation 1:** Table 1 captures which performance metric would be useful for BM-Case1 and BM-Case2 in case of different output options for the AI/ML functionality.**Proposal 1: Option1 (RSRP accuracy) and Option 2 (beam prediction accuracy) are sufficient to cover different cases of AI/ML enabled BM use case for both BM-Case1 and BM-Case2.****Observation 2:** Different options for performance monitoring under discussion in RAN1 BM might potentially have different performance and core requirements impacts in RAN4.**Proposal 2: RAN4 should consider performance and core requirements related to performance monitoring for Option 1 (NW-side performance monitoring) and Option 2 (UE-assisted performance monitoring), if there is an impact on legacy measurement reporting.****Proposal 3: RAN4 should consider impacts on latency requirements for known and unknown TCI states for BM-Case1/BM-Case2 depending upon RAN1 design.****Observation 3:** RAN1 considered system level scenario for evaluation of BM use cases, containing spatial consistency between the beams.**Proposal 4: The Table below should be used as the starting point to discuss various parameters for the testing setup of AI/ML enabled BM use case.**

|  |  |
| --- | --- |
| **Parameters** | **Values** |
| Fading models | AWGN, CDL, TDL |
| Propagation condition | Pathloss, Slow Fading, Fast Fading |
| Maximum number of probes | Legacy or higher |
| Relative angular offset between probes | Same as legacy, multiRx test setup |
| UE rotation | Depending upon if angle of arrival is required in other WGs |
| UE trajectory model | Straight-line trajectory with constant speed |

**Proposal 5: For AI/ML enabled BM use case, a spatial correlation model between the beams, including fading channel model such as CDL channel models needs be considered for the test setup.****Proposal 6: For AI/ML enabled BM use case, RAN4 should target to use same number of maximum probes as in legacy tests.****Proposal 7: Similar approach as was selected in multiRx test setup, should be considered for relative angular offset between probes for AI/ML enabled BM use case. In legacy, relative angular offset between probes of 30°, 60°, 90°, 120° and 150° was considered.****Proposal 8: UE rotation may or may not be required depending upon if AoA is considered in other WGs.****Observation 4:** If an LCM action is required and it is not taken in a timely manner, the performance for AI/ML enabled BM use case may be degraded to undesirable level.**Proposal 9: Core requirements should be considered to limit latency of LCM actions (e.g. activation/de-activation/switching/fallback to legacy) typical for AI/ML enabled BM-Case1 and BM-Case2.****Proposal 10: For the verification/testing of generalization related aspects in RAN4 for AI/ML enabled BM use case, RAN4 should define different scenarios based on parameters listed in the Table below.**

|  |  |
| --- | --- |
| **Parameters** | **Description** |
| Propagation Model | AWGN/CDL/Uma/UMi |
| SINR | Good / Bad Radio conditions |
| UE Speed | Slow / Medium / Fast |
|  Channel propagation conditions | LOS/NLOS |

**Proposal 11: For the verification/testing of scalability related aspects in RAN4 for AI/ML enabled BM use case, RAN4 should define different scenarios based on parameters listed in the Table below.**

|  |  |
| --- | --- |
| **Parameters** | **Description** |
| gNB antenna configurations | gNB antenna array 2x4/4x8/8x16 |
| Variable number of Set B beams | Set B 16/32/64 beams |
| UE Rx beams | UE Rx beams 4/8 beams per panel |

 |

## Open issues summary

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

1. KPIs
2. ideal vs. measured L1-RSRP and impact on BM test
3. Consistency between set A and set B
4. Channel models for tests
5. Test environment limitation/requirements (channels, number of beams, directions, rotation?)
6. Training data set

### Sub-topic 2-1

*Metrics/KPIs for Beam management*

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

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

- Option 1: RSRP accuracy

- Option 2: Beam prediction accuracy

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

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

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

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

- Related measurement accuracy can be considered to determine x

- Option 4: combinations of above options

RAN4 should continue to discuss what metrics are more appropriate and how they impact defining the requirements and testing. There is also a proposal to update Option 3 to also include +x dB.

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

Proposals

* + Option 1: Use Option 1
	+ Option 2: Neither Option 1, 2, 3 is appropriate, a new metric is needed
	+ Option 3: Use Option 2
	+ Option 4: Combination of the above
	+ Option 5: discuss new metrics
	+ Option 6: update Option 3 to also include +x dB as proposed in R4-2404573
* Recommended WF
	+ To be discussed

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

### Sub-topic 2-2

*Measurement accuracy*

Several companies brought up the issue that measurement accuracy will affect the prediction accuracy and how this should be considered.

**Issue 2-2: Measurement accuracy**

* Proposals
	+ Option 1: RAN4 should study the impact of measurement accuracy on prediction accuracy
		- RAN4 should discuss a study framework
	+ Option 2: measured L1-RSRP should be used in the tests, an error model is needed when deriving the requirements
	+ Option 3: requirements should be derived based on ideal measurement results
	+ Option 4: Others
* Recommended WF
	+ To be discussed

Different options can be discussed separately and options are not exclusive

### Sub-topic 2-3

*Consistency between set A and set B*

Multiple contributions brought up the issue about consistency between set A and set B and that this will affect the requirements and tests

**Issue 2-3: set A and set B consistency**

* Issues raised:
	+ Option 1: Consistency(spatial) between set A and set B should be ensured in the test(and in the real deployment
	+ Option 2: Consistency between set A and set B should be taken as a side condition for the requirements
		- FFS how to define the level of consistency
	+ Option 3: RAN4 should consider the size and composition of set B and set A for the prediction accuracy
	+ Option 4: other issues/proposals related to consistency
* Recommended WF
	+ To be discussed

### Sub-topic 2-4

*Channel models in tests*

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

**Issue 2-4: Channel models**

* Proposals
	+ Option 1: CDL channels are needed (with fading)
	+ Option 2: Only use AWGN channels
	+ Option 4: multipath fading is needed, FFS how this can be emulated in the chamber
* Recommended WF
	+ To be discussed

### Sub-topic 2-5

*Test environment limitation/requirements*

Multiple companies brought up several issues regarding the test environment needed. RAN4 should come up with a list of requirements/conditions for the test environment in order to further discuss feasibility and what can be achieved with the current testing setups.

**Issue 2-5: Test environment**

* Proposals
	+ Option 1: Test environment needs to emulate at least X beams for set B. X=8, 16
	+ Option 2: Test environment must ensure spatial consistency between beams in set A and set B
	+ Option 3: number of angles of arrivals (or angle of arrival range/spread)
	+ Option 4: number of AoDs
	+ Option 5: UE rotation
	+ Option 6: others
* Recommended WF
	+ To be discussed

### Sub-topic 2-6

*Training data*

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

* Proposals
	+ Option 1: Training Data set to be specified in RAN4(directly or through some algorithm )
	+ Option 2: Training data set to be left to implementation (companies can generate it based on knowledge of the test environment
	+ Option 3: others
* Recommended WF
	+ To be discussed

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2404428**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404428.zip) | CATT | **Proposal 1: RAN4 to discuss the feasibility and how to define the accuracy requirements for direct positioning.** **Proposal 2: For UE-based AI/ML assisted positioning with UE-side model, RAN4 do not define requirements for the measurement quantities outputted by AI/ML models. The accuracy of inferred quantities and calculated UE positions are guaranteed by the deployed AI/ML models and LCM procedures.** **Proposal 3: Regarding the accuracy requirements for measurements and reported metrics/values, the existing requirements can be a starting point and RAN4 only discuss new requirements if new metrics to be measured / reported are introduced by other groups at current stage.** **Proposal 4: Whether to enhance / tighten the existing requirements can be discussed in performance part based on evaluation results.** **Proposal 5: For UE-sided direct positioning,** **- For inference, positioning accuracy requirements can be considered if UE needs to report.****- For monitoring/test, one method to obtain the ground truth is to request UE to report legacy measurement quantities with the inferred location where the legacy measurement quantities are used to calculate the UE location as ground truth.** **Proposal 6: Quantities in Option 2, i.e., CIR/PDP/channel estimation accuracy, are not appropriate to be metrics/KPIs since they may be the inputs of the AI/ML models.** **Proposal 7: For the quantities in Option 3, i.e., ToA, RSTD and RSRP, and RSRPP, the existing accuracy requirements can be the starting point. Whether to tighten/enhance the existing requirements can be discussed in performance part (Proposal 4).** |
| [**R4-2404572**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404572.zip) | Xiaomi | ***Proposal 1: RAN4 should not define any requirements for case 1.******Proposal 2: RAN4 should define requirements for case 2a/2b.******Proposal 3: RAN4 should define requirements for case 3a/3b at least for accuracy requirements.*****Observation 1: On how to handle the requirements for measurements and reported metrics/values, the current requirement metric (e.g. PRS RSRP) are completely dependent on the reporting with the physical layer reference signal measurements. However, for the direct AI/ML cases (e.g. case 1/2b/3b) in which there are not intermediate measurement results reported from UE but the direct location estimation, currently RAN4 had not any existing requirements reusable.*****Proposal 4: Whether the existing measurement metric could be reused shall be discussed upon RAN1 case by case.******Proposal 5: If RAN1 using sample-based measurement as model input for AI Pos, the existing accuracy requirements shall be restudied (e.g. the margin for the quantization error could be different with that in Rel17).******Proposal 6: AI performance monitoring can be used for the post deployment verification directly.*** ***Proposal 7: RAN4 shall study the necessary and feasible performance hypothesis parameters upon RAN1’s agreement on AI positioning report metric.*** |
| [**R4-2404718**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404718.zip) | CMCC | ***Proposal 1: for case 3a, it is proposed to define positioning accuracy requirements taking legacy performance as baseline.**** ***There are legacy gNB based positioning requirements, i.e. gNB Rx-Tx time difference absolute accuracy in 13.2.2.2 of TS 38.133, and gNB SRS-RSRP in 13.3.2.2 of TS 38.133.***

***Proposal 2: for direct AI/ML positioning, it is proposed to define requirements, and the metrics for positioning requirements/tests is proposed as positioning accuracy: Ground truth vs. Reported.******Proposal 3: except ToA, RSTD and RSRP, and RSRPP, it is proposed to consider CIR/PDP as new measurement to define requirements.******Proposal 4: when specify performance reqirements for AI/ML based positioning, it is proposed to discuss whether and how to consider the impact due to different assumption, e.g.model-input Size Reduction, non-ideal label(s), etc.*** |
| [**R4-2404933**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404933.zip) | Samsung | **Observation 1: RAN4 agreed NOT to define positioning accuracy requirements for case 3a/3b, which are NG-RAN node assisted positioning with network-side model.****Proposal 1: It should be a baseline for future discussion of Rel-19 AI/ML use cases that RAN4 will not consider defining the requirement of the network-side AI/ML models in Rel-19.****Observation 2: RAN4 will come back to case 2a/2b based on progress in the other working groups given that they need more concrete and detail process between UE and LMF to define their requirements.****Proposal 2: Accuracy requirement and report mapping could be reused for case 2a with some enhancements from the legacy requirements while waiting for further detail from RAN1 such as which information is supported for the assistance and how they work in LMF.****Proposal 3: For case 2b, it would be not necessary to consider the requirement of the model inference at the network-side, similar as Case 3a/3b.****Observation 3: For Case 1, the method how to interpret the label data as the prediction performance of the exact UE location would be the key in terms of testing feasibility along with the TE implementation.****Observation 4: Unless RAN4 defines the dataset for TE in advance very well, which could handle the most of scenarios to check the performance benefit, it would be not so meaningful to define such requirement and test method.****Observation 5: Given that RAN4 has not defined the requirement for UE-based positioning in the previous Work Items, it would be very difficult to figure out the performance benefit from the AI/ML based positioning accuracy compared to the legacy non-AI/ML based positioning accuracy.****Proposal 4: RAN4 will not define the positioning accuracy requirements for case 1.** |
| [**R4-2404947**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404947.zip) | vivo | ***Proposal 1: RAN4 to consider the straight-line distance (unit: m) between the real position and the estimated position as the metric for accuracy requirement, if needed to be defined for case 1.******Proposal 2: RAN4 to define performance requirements for the existing timing-related intermediate features: DL RSTD, UE Rx-Tx time difference for case 2a, and FFS whether the already defined requirements can be reused.******Proposal 3: RAN4 to consider define performance requirements for the new metrics, e.g., ToA, if agreed by RAN1 that support to be reported to NW.******Observation 1: It is maybe not possible to define requirement for the LoS/NLoS indicator soft value.******Proposal 4: RAN4 to define performance requirement for LoS/NLoS indicator hard value, the accuracy requirement can be the probability of maximum LoS/NLoS misestimation.******Proposal 5: RAN4 to wait RAN1 progress to decide whether to define requirement for phase information.******Proposal 6: RAN4 to define requirement for delay and power for case 2b, whether the measurement is sample-based measurement or/and path-based measurement is based on the progress of RAN1.******Proposal 7: RAN4 to consider define requirement for power/delay information as the vector, e.g., ΔRSRP=SGCS(RSRPreal, RSRPestimated), where RSRPreal= [RSRPreal1, RSRPreal2, …, RSRPrealN] is the real power information, RSRPestimated= [RSRP1, RSRP2, …, RSRPN] is the estimated power information, N is number of sample/path.******Observation 2: PRUs can be deployed in the test scenario to provide measurements together with the location information.******Proposal 8: RAN4 to consider deploy PRUs in pre-deployment test so as to derive the ground truth value of UE position and intermediate features.******Observation 3: At least LoS/NLoS indicator hard value can be tested since the LoS/NLoS environment is known in test system.******Proposal 9: RAN4 to discuss the key factors that affect the performance and consider introduce reference scenario/model/dataset for AI/ML based positioning test.*** |
| [**R4-2405150**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405150.zip) | Huawei,HiSilicon | ***Proposal 1:*** For UE-sided direct AI/ML positioning (Case 1), positioning accuracy is not testable, if the positioning result derived by UE is not reported to NW.***Observation 1:*** For UE-assisted positioning with LMF-side positioning (Case 2b), the relationship between measurement accuracy and positioning accuracy is unavailable, which has an impact on the test requirement definition.***Proposal 2:*** If RAN4 studies how to test Case 2b in AIML for positioning, the relationship between measurement accuracy provided by UE and the eventual positioning accuracy at LMF needs to be investigated firstly. |
| [**R4-2405186**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405186.zip) | OPPO | **Proposal 1: For case1, RAN4 will not define positioning accuracy requirements in R19 WI.****Proposal 2: For case1(AI/ML assisted positioning), FFS the test feasibility, FFS 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.****Proposal 3: For case2a/2b, should be treated with 2nd priority in RAN4 R19 WI.****Proposal 4: For case3a/3b (cases without UE-side model), not necessary to test the Positioning model/functionality outputs.** |
| [**R4-2405189**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405189.zip) | ZTE Corporation | **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, that is, it only needs to determine what the input is, how to calculate it, and how the investment is the network implementation behavior.****Observation 2: For the AI/ML positioning, the ground truth values are hard to obtain and we know the location of PRU, the PRU can be the reference UE location for AI positioning.****Proposal 1: PRU can be the reference location for AI/ML positioning.****Observation 3：RSTD, RSRP and RSRPP these accuracy requirements have already been defined in legacy.****Proposal 2: RAN4 to study how to define test and accuracy requirements for ToA.****Proposal 3: The discussion of LOS/NLOS indicator shall be deprioritized since RAN4 has never defined the requirements or test for indicator.** **Proposal 4: RAN4 shall study how to define requirements and tests for model input including new measurements such as CIR/PDP.** |
| [**R4-2405520**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405520.zip) | Ericsson | **Proposal 1**: RAN4 to not define requirements (core and performance) for use case 1.**Proposal 2**: For use cases 2a/2b, RAN4 can study the potential requirements, but RAN4 will define requirements for use cases 2a/2b only after some progress has been made in RAN2. |
| [**R4-2405708**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405708.zip) | Apple | **Proposal 1: RAN4 to further discuss the feasibility and how to define requirements for Positioning accuracy for case 1. Positioning test data sets could be one option for testing this KPI.** **Proposal 2: RAN4 to define performance requirements for use case 2a. The specifics and scope of the core and performance requirements for use case 2a depend on the outcomes of discussions within RAN1 and RAN2****Proposal 3: For Assisted AIML Positioning, the KPIs test metric (e.g., LOS/NLOS) needs to be considered for validating the positioning accuracy****Proposal 4 : In AI/ML-based positioning, it is essential to investigate the performance requirements for the input parameters of the positioning model/functionality (e.g., measurement accuracy of RSRP, ToA, RSRPP, RSTD) across all AI/ML positioning cases****Proposal 5: RAN4 to study defining performance accuracy requirements for use case 2b. The specifics and scope of both core and performance requirements for this use case depend on the outcomes of discussions within RAN1 and RAN2****Proposal 6: 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 7: RAN4 to define performance accuracy requirements for measurments 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** |
| [**R4-2405810**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405810.zip) | Nokia | 1. If a new measurement is supported (CIR, DP, and PDP), it may need to define a new requirement.

If a legacy measurement is supported (RSTD, RSRP, and RSRPP), it may need to study whether to enhance or reuse a new requirement.RAN4 should define the measurement accuracy requirements if RAN1 supports any new measurement(s) as AI/ML functionality input in Case 1. Positioning accuracy should be considered as the performance metric in Case 1 (UE-based positioning with UE-side model, direct AI/ML positioning).Positioning accuracy can be verified based on the ground truth which may consist of the location points with known positioning co-ordinates (e.g., PRU or GNSS based).RAN4 should further study the feasibility of test mechanisms for positioning accuracy metric verification.RAN4 should further study feasibility of using synthetic channels for testing in Case 1. RAN4 should consider to modify and re-use the legacy test mechanism described for A-GNSS test in 37.571-1 clause 13.6 for validation of Positioning accuracy KPI for Direct Positioning.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.RAN4 to define the time latency limit on UE’s LCM actions when an LCM procedure is indicated by network. 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).RAN4 should further study the feasibility of the test mechanisms for LOS/ NLOS metric verification for case 2a (UE-assisted/LMF-based positioning with UE-side model, AI/ML-assisted positioning).RAN4 to discuss whether any accuracy measurement requirement is needed for at least the types of time domain channel measurements supported by RAN1 in Case 2b.  |

## Open issues summary

In the previous meeting it was already agreed that RAN4 will not define positioning accuracy requirements for case 3a/3b. The discussion will continue for the other topics.

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

1. Requirements for case 1
2. KPIs for case 1
3. LOS/NLOS indicator

### Sub-topic 3-1

*Requirements for case 1*

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

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

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

To be discussed

Note: currently there are no requirements for UE based positioning

### Sub-topic 3-2

*KPIs for case 1*

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

Identified KPIs in the TR:

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

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

**Issue 3-2: KPIs for case 1**

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

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

### Sub-topic 3-3

*LOS/NLOS indicator*

There are some proposals to introduce requirements/tests for the LOS/NLOS indicator

**Issue 3-3: LOS/NLOS indicator**

* Proposals
	+ Option 1: RAN4 should discuss requirements for the LOS/NLOS indicator
	+ Option 2: RAN4 should wait for progress in other groups whether LOS/NLOS reporting is defined
	+ Option 3: It is not feasible for RAN4 to define requirements for such an indicator
	+ Option 4: others
* Recommended WF

To be discussed

### Sub-topic 3-4

*KPIs for case 3a/3b*

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

Identified KPIs in the TR:

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

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

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

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

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

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

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

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| [**R4-2404153**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404153.zip) | MediaTek inc. | **Proposal 1**: RAN4 should clarify what is the meaning of “fully specify” for test decoder Option 3. * Option 1: Fully specify the weights/bias for each neuron in a neutral network.
* Option 2: Specify model architecture and training related parameters.

**Proposal 2**: When discussing parameters to describe test decoder Option 3, training data and hyperparameters should be taken into consideration as they may influence the performance.**Observation 1**: RAN1 is now also discussing standardize model/dataset/dataset format to resolve the inter-vendor interoperability issue.**Proposal 3**: Postpone the discussion for the Option 3 and 4 and wait for RAN1’s conclusion on inter-vendor interoperability issue. |
| [**R4-2404206**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404206.zip) | VIAVI Solutions | **Observation 1:** Type 1 training collaboration does not permit model confidentiality and should therefore not be considered for RAN4 tests.**Observation 2:** Type 2 simultaneous training would considerably slow down testing and not scale.**Observation 3:** Type 3 collaboration with NW / TE side first training would allow each TE vendor to then share a set of information (e.g. dataset) for all UE / DUT vendors to use for training their encoder models (UE side CSI generation part) ready for testing.**Observation 4:** The performance loss for type 4 training collaboration relative to 1-on-1 joint training can be similar to that of type 2 with the right special handling (e.g. an adaptation layer).**Proposal 1:** Adopt a type 3 training collaboration approach for RAN4 testing and consider special handling (e.g. an adaptation layer) to minimise performance variations between TE vendors. |
| [**R4-2404216**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404216.zip) | Qualcomm Incorporated | **Observation 1: Some of the layer/model sizes are determined by the test configurations, and therefore we can parameterize the layer sizes as a function of test configuration parameters, add rows in the table for test configurations, to simplify the discussion. Therefore, the agreed table can be applicable to multiple test configurations if needed.****Proposal 1: We propose a set of test decoder model parameters in Table 1 and depict it in Figure 1. Besides the test decoder , we propose to add a new category: test configuration (rows are in orange). The rows in the test configuration list some parameters used in the layer size and model type to separate the test configuration discussion from the model design discussion.** **Observation 2: If model architecture parameters are agreed, different training procedures and hyperparameters are likely to lead to very similar decoders.** **Propose 2: Remove training procedure from “the parameter needed to be agreed”, RAN4 shouldn’t discuss training procedures, hyperparameters and cross-validation details before model architecture parameters are agreed.** **Proposal 3: RAN4 consider the following options for deriving the partially specified test decoder specifications:*** **Option 4a-1 standardized dataset**
	+ **Step 1: RAN4 agrees a pair of encoder and decoder with full details (same as fully specified decoder discussion) and an encoder input data generation procedure.**
	+ **Step 2: RAN4 uses this encoder/decoder pair and the generation procedure to generate a set of decoder input and output data and captures this dataset in the specification.**
	+ **Step 3: RAN4 specifies a test decoder verification procedure based on the specified dataset.**
* **Option 4a-2 standardized aggregated dataset**
	+ **Step 1: RAN4 achieves some agreements (e.g., part of but not all the parameters in the test decoder parameter table in the previous meeting WF[1]) for the test decoder.**
	+ **Step 2: Interested companies can design their own encoder/decoder pairs based on the agreements to contribute the (decoder input, decoder output) dataset to RAN4**
	+ **Step 3: RAN4 aggregates the datasets from all the contributing companies, and capture the aggregated dataset in the specification**
	+ **Step 4: RAN4 specifies a test decoder verification procedure based on the specified dataset.**
* **Option 4b reference encoder**
	+ **Step 1: RAN4 agrees a pair of encoder and decoder with full details (same as fully specified decoder discussion) and an encoder input data generation procedure**
	+ **Step 2: RAN4 capture the encoder as a reference encoder and the encoder input data generation procedure in the specification.**
	+ **Step 3: RAN4 specifies a test decoder verification procedure based on the reference encoder.**

**Proposal 4: Feasibility and concerns of the partially specified test decoder option should be analysed from the perspective of test repeatability (decoder output consistency/similarity given the same encoder output (latent message) among test decoders implemented by TE vendors). The following feasibility issues and concerns for options 4a-1, 4a-2 and 4b (as described in proposal 3) need to be resolved:*** **Option 4a-1 (standardized dataset from one enc/dec pair): whether it is feasible to design a standardized dataset sufficiently representing the propagation channel condition and possible UE procedures (channel estimation, whitening and desired precoder derivation). If the dataset doesn’t cover all the possible UE procedures, the encoder input not captured in the standardized dataset can produce very different decoder output, which violates the consistency requirement in option 4.**
* **Option 4a-2 (standardized dataset from multiple enc/dec pairs): whether it is feasible to design a test decoder that can properly recover the encoder input from the encoder output, given that the dataset is from multiple encoder and decoder pairs**.
* **Option 4b (reference encoder): required to define a proper encoder input data generation procedure and a reference encoder with a good coverage of latent space to guarantee decoder output consistency by the verification procedure with the reference encoder. Another concern is that this option implies a more complicated procedure to train UE encoder, compared to option 3 fully specified decoder.**

**Observation 3: RAN4 can consider Table 1 and Figure 1 for derivation of the highlighted entries in the below table (copy of Table 2) if option 4a-1 or 4b is chosen.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Option 3 fully specified decoder | Option 4a-1 standardized dataset | Option 4b reference encoder | Option 4a-2 standardized aggregated dataset |
| Required RAN4 agreement (in the WF, not spec) | * Encoder input generation procedure
* (Likely to have it as the by-product of test decoder) an encoder with full details
 | * An encoder/ decoder pair with full details
* Encoder input generation procedure
 | * Encoder input generation procedure
* (Likely to have it as the by-product of reference encoder) a decoder with full details
 | Part of the encoder/decoder model parameters  |
| RAN4 specification | A test decoder with full details | * Dataset of (decoder input, decoder output) from the agreed encoder/ decoder pair
* Decoder verification procedure
 | * A reference encoder with full details
* Decoder verification procedure
 | * Aggregated dataset of (decoder input, decoder output) from each contributing companies’ encoder/ decoder pair
* Decoder verification procedure
 |
| Note that “with full details” refers to agreed parameters from model structures, depth, size, quantization etc, i.e., every agreed parameters in the test decoder parameter table in the previous meeting WF[1]. |

**Proposal 5: RAN4 continues to study a feasible (from test repeatability perspective) and implementation/specification friendly solution to option 4.** |
| [**R4-2404429**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404429.zip) | CATT | **Observation 1: RAN4 need to define a specific value/threshold if absolute throughput is adopted.** **Observation 2: RAN4 need to define a baseline if absolute throughput is adopted, and the existing baseline is derived based on random PMI.** **Proposal 1: RAN4 to clarify the baseline first before deciding which type of throughput is adopted in AI/ML tests. The throughput derived based on the non-AI/ML method can be the baseline.** **Observation 3: Different model types are adopted by RAN1 for different use case simulations. And the model structures can be completely different for different model types.** **Proposal 2: RAN4 decide model type case by case and then discuss other model parameters. The transformer models illustrated in Figure 2.4-3 and 2.4-4 in R4-2401612 can be a starting point for CSI compression.** **Proposal 3: The following parameters should be decided after the model type is determined:** **- 1st priority: Model depth / layer size / Layer type / training datasets / loss function, etc.** **- 2nd priority: Training procedure / quantization method for the encoder output, etc.** **- 3rd priority: Depends on the evaluations.** **Proposal 4: Constraint conditions should be defined for the test decoder being evaluated after model type is decided. The possible constraint conditions are as follows and companies can provide more candidate options:** **- The maximum memory (e.g., number of parameters);****- The maximum model complexity (e.g., in FLOPs);****- Evaluation metrics (absolute / relative throughput, etc.)****- Other constraint conditions.** **Proposal 5: After evaluations, select one trained model among the models provided by all companies based on some criteria which can be FFS. This selected model will be fully specified in RAN4 spec.**  |
| [**R4-2404430**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404430.zip) | SEU | **Observation 1: CsiNet has been extensively studied and a large number of derivative models have emerged based on it****Observation 2: Sufficient evaluation results can be provided for CsiNet and its derivative models with their source code.****Proposal 1: CsiNet can be considered as a candidate backbone structure for AI-based CSI compression with a two-side model** |
| [**R4-2404477**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404477.zip) | CAICT | **Proposal 1: Beside throughput, other candidate options (e.g., applicable condition, dataset characteristic and etc.) could be considered for discussion on monitoring metrics.****Proposal 2: Suggest to agree “Encoder-decoder interface”, “Fixed point representation”, “Format of input to encoder/output of decoder” and “Hyperparameters” in the table.****Proposal 3: Clarifications are required for “Number of bits of latent message”:*** **Q1) “*Number of bits of latent message*” would be a specific value or a value range?**
* **Q2) Do we need specify multiple “*Number of bits of latent message*”?**
* **Q3) If yes for Q2, does a specific “*Number of bits of latent message*” couple with specific testing configurations and/or other parameters in the table?**

**Proposal 4: Explicit assumptions should be specified for building training dataset of option 3, according to targeting scenarios/configuration/conditions we want to test.****Proposal 5: Certain limitation(s) on model complexity could be considered for option 3.****Proposal 6: Suggest to discuss whether a single testing decoder or multiple decoders shall be defined for CSI compression.****Proposal 7: Suggest to discuss the proposed outline of following working procedure for option 3.*** **Step 1: for fair comparison, we may need to specify a common dataset, e.g., by merging datasets from difference vendors, or selecting a reference dataset from Third Party platform.**
* **Step 2: Define evaluation methodology, especially the performance metrics and evaluate different models based on EVM.**
* **Step 3: Decide the model type and proper model structure (e.g., layer number, layer size) based on evaluation results.**
* **Step 4: Finalize detailed parameters of the model.**
 |
| [**R4-2404574**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404574.zip) | Xiaomi | **Proposal 1: RAN4 to discuss the assumption for reference decoder first, e.g. whether time domain compression or prediction is considered or not, which will have impact on the reference model type and structure.****Proposal 2: RAN4 needs to discuss how to define a rule to choose a reference decoder model if model architecture and parameters from companies are diverse.** |
| [**R4-2404720**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404720.zip) | CMCC | ***Proposal 1: for CSI Prediction Accuracy metrics for inference, it is propose to use relative throughput, which is the throughput gain achieved with predicted PMI compared to random PMI.*** ***Proposal 2: for CSI compression and CSI prediction, it is proposed to use intermediate KPI, e.g. SGCS, as requirements/tests metrics for LCM.*** |
| [**R4-2404932**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404932.zip) | Intel Corporation | **Proposal #1: RAN4 should further align its test methodology with RAN1's conclusions on inter-vendor training collaboration for AI/ML-based CSI compression.****Proposal #2: Further discuss and define the upper bound complexity including the number of computations and number of parameters for Option 3 and Option 4 reference decoders.****Proposal #3: Further agree on the steps required to conduct analysis of Option 3 and on the principles to select the decoder for specification.****Proposal #4: Request RAN1 inputs on the identified decoder models for Option 3 and 4 taking into account performance/complexity.****Proposal #5: Conduct Option 4 analysis in parallel with Option 3 reusing same assumptions on Model architecture parameters and Training data set for model training.** |
| [**R4-2404948**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2404948.zip) | vivo | **Observation 1: Model structure (back-bone, parameters, e.g., number of layers, etc) also has significant performance impact even if complexity of model (in terms of FLOPS) are similar.** **Observation 2: RAN4 testability study should consider all the relevant parts for defining performance requirements and testing.****Proposal 1: RAN4 to define reference model for defining performance requirements for one-sided model.****Proposal 2: In 2-side model use case, both reference encoder and reference decoder are introduced for defining performance requirements for UE side encoder.****Proposal 3: Fully specified and partially specified options, i.e., option 3 and/or option 4, are used as baseline for RAN4 to specify reference model for defining requirements for different use cases for both 1-sided model and 2-sided model.****Proposal 4: There would be three potential sub-options for Option 4:*** **Option 4a: Specify encoder.**
* **Option 4b: Specify model structure of decoder.**
* **Option 4c: Specify model structure of encoder.**
* **Note: Dataset would need to be aligned or specified in Option 4, where the data content may be different in different sub-options.**

**Proposal 5: Option 3 and Option 4 can work together. The encoder and the dataset should be aligned when specifying the decoder.** **Proposal 6: The reference/test encoder/decoder may be aligned through the following procedures*** **Step 1: Align the dataset containing only channel information.**
* **Step 2: Determine the model hyperparameters that need to be aligned.**
* **Step 3: Define the evaluation method for model complexity and performance.**
* **Step 4: The best model structure(s) may be selected based on the aligned evaluation method, through the simulations using the aligned dataset.**
* **Step 5: Based on the aligned model structure, the specific parameters of the reference model would be merged from companies**

**Proposal 7: Take into consideration the parameters to be aligned for Option 3 and Option 4 in Table 2.4-1.**

|  |  |  |
| --- | --- | --- |
| **Category** | **Parameter** | **Description/Examples** |
| Model architecture parameters | Model type | Transformer, CNN, RNN, MLP |
| Model depth | Number of layers |
| Layer type | Fully connected, convolutional, activation layer, normalization layers, etc. |
| Layer size | Neuron count and configuration |
| Quantization method for the encoder output | Scalar, vector (with codebook) |
| Encoder-decoder interface | Number of bits of latent message |
| Fixed point representation | Int8, int16, floating point etc |
| Format of input to encoder/output of decoder |  |
| Special connection relationships between layers (if needed) | ResBlock, Inception, etc |
| Special hyperparameters for CNN (if needed) | Kernel size, stride, padding, activation function, bias, and channel number. |
| Special hyperparameters for Transformer (if needed) | Implementation method for multi-head attention, parameters for multi-head attention such as number of heads and dimensions of heads. |
| Model Training related parameters | Training procedure | FFS (e.g Initialization method, training duration, training completion criteria, collaboration type, encoder assumption, etc)Note that training procedure does not need to be fully aligned. |
| Loss function | SGCS, NMSE, etc. |
| Training datasets | Channel model, number of Tx/Rx portsOther parameters FFS (e.g. rank)Dataset containing only channel information, which is merged by data from companies. |
| Hyperparameters | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc.Note that training procedure does not need to be fully aligned. |
| Cross-validation details | Dataset splits for training/testing/validation |
| ~~Generalization (may be applicable to all four options)~~ | ~~Performance requirements on test dataset(s)~~ | ~~Mean SGCS, etc.~~ |
| Scalability ~~(may be applicable to all four options)~~ | Supported antenna port configurations | (2,8,2), (2,4,2), etc. |
| Supported feedback payloads | Low, medium, high overhead (with specified number of bits) |
| The evaluation method | The evaluation of models considering both model complexity and performance |  |

**Proposal 8: The suggested model structures in Figure 2.4-1 to 2.4-4 for test decoder/encoder could be used as a starting point****Proposal 9: “Supported training collaboration type between DUT and decoder provider” can be removed from the table of the comparison of the four options of test decoder, since this aspect is just for training before test and seems to have no obvious impact on the test.****Proposal 10: Take into consideration the summary of 4 options for testing of 2-sided model in Table 2.4-2.****Proposal 11:** **Compared with absolute throughput, relative throughput would be used to see the gain from CSI prediction. The comparison baseline can be further discussed, such as randomly chosen PMI.****Proposal 12: Since Monitoring is still under discussion in RAN1, RAN4 should wait for further progress of RAN1.** |
| [**R4-2405151**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405151.zip) | Huawei,HiSilicon |  ***Proposal 1***: Deprioritize Options 1 and 2 for determining the test decoder of two-sided model.***Proposal 2***: According to whether using a mixed training dataset to determine the reference decoder, Option 3 can be further divided into two sub options as follows.* Option 3a: The test decoder is determined for each test case, using a specific dataset collecting from the configuration/scenario of the considered test case.
* Option 3b: The test decoder is determined for more than one test cases, using a mixed dataset collecting from different configurations/scenarios of the considered test cases.

***Proposal 3***: Take Option 3a as baseline, where a specific rather than a mixed dataset is used for defining the test decoder in each test case. ***Proposal 4***: For achieving a converged test decoder in Option 3, at least the structure of both the reference encoder and test decoder, hyperparameters of model training, as well as a determined sample-by-sample dataset are expected to be aligned among companies. ***Observation 1***: Even with all hyperparameters aligned and model training converged, the model parameters provided by companies can still be different. How to align model parameters of the test decoder among companies is an open issue.***Proposal 5***: According to whether the model structure is specified, Option 4 can be further divided into two sub options as follows.* Option 4a: Model structure is not specified in RAN4. Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder.
* Option 4b: Model structure is specified in RAN4. Training dataset is not specified for verifying the encoder at DUT. The test decoder developed by TE vendor needs verification.
	+ FFS: How to determine the test metric for test decoder developed by each TE vendor.

***Observation 2***: The boundary between Option 3 and Option 4b is whether the model parameters are specified in RAN4.***Proposal 6***: In Option 4, the performance of test decoder should be verified before testing DUT in each test.***Proposal 7***: Compression ratio and quantization level needs to be specified in Options 3 and 4.***Observation 3***: Though there is no need to align the model parameters of test decoder, model parameters of the reference encoder for verifying the test decoder is still needed. How to align is still an open issue.***Proposal 8***: The comparison of the four options of test decoder is updated as below. **See paper*****Proposal 9***: Deprioritize SCGS/NMSE for defining baseline requirements in AIML-enabled CSI compression. ***Proposal 10***: Deprioritize CSI prediction accuracy for defining baseline requirements in AIML-enabled CSI prediction. ***Observation 4***: How to ensure that the testing dataset aligns well with training dataset is still an open issue. |
| [**R4-2405187**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405187.zip) | OPPO | **Proposal 1: Principles to define test decoder(s)*** **to meet the minimum performance requirement in RAN4 tests**
* **to be a simple design**

**Proposal 2: Steps to determine test model(s):*** **Setp1: Determine the test condition**
* **Step2: Determine the test data, including**

 **Step 2-1: consider how to set up the test data, how to determine the test data generation method**  **Step 2-2: consider how to collect datasets from different companies** **Step 2-3: consider whether/how to merge datasets from different companies to form a reference model data set*** **Step3: Determine the test model structure**

 **Step 3-1: consider how to set up the model structure** **Step 3-3: consider how to merge and compromise different inputs on the proposed reference model structures from different companies, to form a reference model structure*** **Step4: Determine the test model parameter**

 **Step 4-1: determine the reference model parameters based on the reference dataset and reference model structure** **Step 4-2: consider how to merge different reference model parameters to obtain the specified reference model parameters****Proposal 3: For both CSI compression and CSI prediction, model/functionality input (CSI-RS measurement accuracy) and output (associated throughput) related tests should be supported.****Proposal 4: Existed RAN4 test examples for “reporting of PMI” can be reused or serve as a reference. Requirement of γ and test settings can be reused or updated.****Proposal 5: Reuse the legacy PMI requirement (compare to random precoding) as a baseline test, and let other options/proposals with higher performance requirements be further studied.****Proposal 6: In R19, static test scenarios and configurations should be considered first. After having feasible testing cases for static configurations, then further consider whether to introduce non-static testing scenarios and configurations.****Proposal 7: Besides the TDL channel based tests(could be a baseline test), CSI-related tests can be conducted under CDL channel [or other more practical channel conditions] to check a relatively generalized performance.** |
| [**R4-2405212**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405212.zip) | ZTE Corporation | ***Observation 1. According to our analysis, we can observed that two different candidate options for channel matrix, which have different input formats to encoder.******Observation 2. To reduce feedback, channel information could be sparse in some domains, which will lead different input formats to encoder.******Observation 3. For test decoder, need more information about training dataset and model structure in compression.******Proposal 1. For CSI prediction, it is proposed to use relative throughput as default metric.******Proposal 2. Format of input to encoder/output of decoder could be considered, which also including input dataset pre-processing.*** ***Proposal 3. Our views on two-sided testing options comparison are listed in the following table with additional notes.*** |
| [**R4-2405611**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405611.zip) | Ericsson | [Observation 1 The RAN1 “reference decoder” and RAN4 “test decoder” are very closely linked](#_Toc163476027)[Observation 2 Eventually there will be a need to decide whether to capture standardized encoder/decoders in RAN1 or RAN4 specifications.](#_Toc163476028)[Observation 3 The RAN1 and RAN4 options have similarities but are not directly comparable.](#_Toc163476029)[Observation 4 To create a proposed test decoder, it is necessary to also consider the structure of an encoder.](#_Toc163476030)[Observation 5 Pre-processing using enhanced type 2 PMI can reduce the model complexity.](#_Toc163476031)Based on the discussion in the previous sections we propose the following:[Proposal 1 The standardized decoder/encoder options for ensuring interoperability are discussed in one WG](#_Toc163476032)[Proposal 2 Align the proposals for standardized encoder/decoder (parameters) between RAN1 and RAN4.](#_Toc163476033)[Proposal 3 For option 3, in addition to a standardized test decoder consider capturing a standardized reference encoder in the specifications.](#_Toc163476034)[Proposal 4 RAN4 should discuss encoder model structure, and the extent to which the encoder structure might differ from the decoder.](#_Toc163476035)[Proposal 5 RAN4 should decide and agree on whether the target is purely performance or both performance and complexity. (Some checking with RAN1 may be needed).](#_Toc163476036)[Proposal 6 Agreements on the scope of the testing (for example channel models, range of Doppler) are needed.](#_Toc163476037)[Proposal 7 For option 4, standardize a reference encoder and a test data framework and test dataset/channel model.](#_Toc163476038) |
| [**R4-2405653**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405653.zip) | Samsung | *AI-CSI prediction***Observation 1: RAN4 agreement on CSI prediction accuracy metrics, i.e., throughput as default metrics, is related to specifying RAN4 requirement only if the sub-use case of CSI prediction is confirmed (based on RAN1 study and decided in the checkpoints in RAN#105 (Sept ’24)).****Observation 2: No matter relative or absolute throughput metrics are adopted for CSI prediction accuracy metric in RAN4 requirement, it is difficult to identify whether AI/ML or non-AI/ML based CSI prediction is used in UE-sided CSI prediction.****Proposal 1: Absolute throughput shall be adopted as the RAN4 performance metric for CSI prediction accuracy requirement for UE-sided AI/ML-based CSI prediction.****Observation 3: For the generalization/scalability performance of AI/ML CSI prediction, RAN1 is still planned to study the AI/ML model performance over various configurations including:** **• Various UE speeds (e.g., 10km/h, 30km/h, 60km/h, 120km/h)****• Various deployment scenarios****• Various carrier frequencies (e.g., 2GHz, 3.5GHz)****• Various frequency granularity assumptions****• Various antenna port numbers (e.g., 32 ports, 16 ports)****Proposal 2: No need further study on the metrics or other aspects for the model monitoring for CSI prediction, until other WGs finalize the corresponding procedure.***AI-CSI compression:***Observation 4: The RAN1 Option 1 for inter-vendor training collaboration, i.e. fully standardized reference model (structure + parameters), is the same as RAN4 Test decoder Option 3, in terms of standardization effort to specify reference model (or called test decoder in RAN4).** **Proposal 3: For RAN4 Test decoder Option 3 (similar as RAN1 Option 1 for fully standardized reference model), RAN4 continue to study the feasibility in terms of standardization procedure.** **Proposal 4: The procedure shown in Figure 1 shall be followed to fully specify the test decoder for AI/ML-based CSI compression and accordingly RAN4 performance requirement:** **Step-1: Identify necessary Model Architecture Parameters****Standardization Procedure End**for a certain use case(e.g., CSI compression for precoding matrix under certain config.)Model architecture parameters could include: Model type, Model depth, Layer type/size, Quantization, etc. Model training procedure, loss function, training datasets, hyperparameters, etc.**Step-3: Companies provide two-sided model design based on their own study/preference****Step-4: Performance comparison based on different companies’ en/decoder designs****Yes****No****Step-6: Performance alignment by companies based on agreed model architecture/training parameters**Performance comparison in terms of metrics like NMSE, SGCS, etc.**No****Yes****No****Standardization Procedure Start****Step-2: Identify necessary Model training Parameters**Test decoder is expected to be generated in this stepReference encoder is assumed, but leave enough implementation flexibility to vendors (similar to Demod alignment for MMSE-IRC)**Step-8: performance alignment for encoder design by companies based on assumptions on reference encoder** **Yes**RAN4 performance requirement obtained (for certain reference encoder)**Step-10: Derive RAN4 performance requirement****Step-5: RAN4agree on two-sided model architecture / training parameters?****Step-7: RAN4 agree on test decoder (which can be fully specified in spec.)****Step-9: RAN4 achieve performance alignment?****Figure 1. Standardization procedure to fully specify test decoder and RAN4 performance requirement****Observation 5: As long as training data set (including enough amount of data for raw CSI and compressed bit strings) available in 3GPP standard for TE decoder training, it is possible for TE vendor and DUT vendor to develop test decoder and interoperable encoder respectively.****Observation 6: RAN1 Option 2 for inter-vendor training collaboration (i.e., standardized dataset) is one possibility of RAN4 test decoder option 4 (defined as TE vendor developed based on standard).****Observation 7: If the training data set (including enough amount of data for raw CSI and compressed bit strings) is available in 3GPP standard, Option 4 can be regarded as the standardized training data set.****Proposal 5: For test decoder Option 4,** **- It is assumed that TE vendor will not share decoder to other vendors (DUT and/or infra vendors);****- Parameters that need to be specified for defining test decoder shall include:** **🡺 Training data set for TE decoder training, including enough amount of data for raw CSI and compressed bit string.** **Proposal 6: RAN4 adopt the following text proposal (yellow-highlighted) for test decoder Option 1-4 in test decoder for 2-sided model Table 7.3.2.3-1 in TR 38.843.**  |
| [**R4-2405709**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405709.zip) | Apple | **Observation 1: There are numerous critical issues that must be addressed to assess the feasibility of option 3.****Observation 2: Option 4b-1 faces similar issues to Option 4a-1, where variations in UE encoder implementations (DUTs) can challenge the behaviour of the test decoder. This is an issue when the decoder has been verified using encoder inputs solely from a reference encoder.****Observation 3: The training/testing procedures that involve dataset A (test decoder training data), dataset B (DUT training data: in-site TE or outside), dataset C (testing data for DUT + Test decoder), and dataset D ( inference for DUT + NW decoder with real field data) should be investigated together to conclude on the overall testing goal feasibility for both options 3 and 4** **Observation 4:** **Multiple standardized training datasets are necessary to cover the identified scenarios and configurations for testing requirements. This would imply the necessity of multiple test decoders. (same model, but each test decoder has different parameters)****Proposal 1: When specifying the test decoder for option 3, RAN4 to consider including regularization techniques, such as weight constraints, dropout layers and their associated probabilities, early stopping criteria, noise injection, and other relevant factors. The size of the training data should also be specified. We updated the table by incorporating our additional proposals for supplementary parameters to specify the test decoder.****Proposal 2: When specifying the test decoder for option 3, RAN4 to consider the options for the training type in the training procedure of the test decoder. Consider training type 1 and 3 as described in this section.****Proposal 3: To assess the feasibility of option 3, RAN4 should address several key considerations. These include:** * **With training type 1 it would be unrealistic to expect the test decoder to produce meaningful answers for all expected DUT UEs**
* **Whether the information provided in the specification to implement the test decoders is adequate to ensure reliable performance across infra-vendors in real deployment (NW vendor implementation should be constrained to match test decoder implementation, but it is very likely that each NW vendor has a proprietary decoder architecture)**

**Proposal 4: For option 3 of specifying the test decoder, prioritize training type 3 using a collaborative dataset sourced from different UE vendors to enhance the generalizability of the test decoder for testing multiple vendors. Dataset can be stored in a repository.** **Proposal 5: Investigate the feasibility of achieving bounded variations across different test decoder implementations with the same UE DUT, considering varying encoder output characteristics resulting from different encoder implementations and channel estimation procedures for Option 4a-1 test decoder implementation.** **Proposal 6: Investigate the feasibility of implementing a test decoder capable of learning the mapping between different encoder inputs and encoder outputs, with the diverse dataset spanning different encoder and decoder implementations for option 4a-2. The training dataset generated by multiple encoder/decoder pairs could be stored in a repository.****Proposal 7: We propose to consider options 3 (full specified test decoder) with training type 3, and option 4a-2 for partially specified test-decoders. For option 4, the objective should be to identify the optimal balance that enables the ecosystem to leverage the advantages of a standardized decoder, while simultaneously preserving opportunities for decoder model distinctions wherever feasible.****Proposal 8: RAN4 to investigate the tradeoffs of training the DUT offline inside the TE or in the field with real-world data, or through a combination of both approaches. Multiple training datasets scenarios associated with a training ID could be linked with corresponding models. This set of models can be integrated into the UE capability and become part of the tests. During testing procedures, assistance information can be utilized to select the appropriate DUT model along with the appropriate test decoder (set of parameters for identified scenario/configuration)****Proposal 9: 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 (DUT training takes place outside the TE).** **Proposal 10: 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**
 |
| [**R4-2405738**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405738.zip) | Nokia | **On Metrics/KPIS for CSI feedback:****Observation 1**: Type I single-panel codebook is easier to randomize to establish the reference PDSCH throughput (in the denominator of relative throughput γ) in comparison with Type II codebook or AI/ML-based compressed feedback.**Proposal 1: Random precoding matrices from Type I single-panel codebook (mode 1 with appropriate number of layers) can be used by SS for evaluation of PDSCH throughput t\_rnd in the denominator of relative throughput γ.****Proposal 2: For AI/ML-based compressed CSI feedback, decoded eigenvectors of the channel matrix reported by the UE can be used for PDSCH precoding at SS for the evaluation of throughput t\_ue in the numerator of relative throughput γ. Details can be further clarified when the feedback is fully specified in the other WGs.****Observation 2**: We are expecting that in CSI prediction use-case legacy CSI reporting format will be used.**Proposal 3:** For AI/ML-based CSI prediction, CSI feedback predicted and reported by the UE in legacy format (e.g., Type II) can be used by SS for PDSCH precoding for the evaluation throughput t\_ue in the numerator of relative throughput γ, i.e., like in legacy requirements.**Observation 3**: The use of the same/comparable metric (relative throughput γ) for AI/ML based and existing legacy requirements with the same parameters gives the way to compare the minimal level of performance.**Proposal 4: RAN4 to ensure that relative throughput based on AI/ML CSI feedback with the same overhead is at least not worse than in legacy tests with the same parameters.****Observation 4**: For CSI prediction performance monitoring, RAN1 is discussing NW-side mechanism based on UE-reported ground truth. This approach can be applied to test the CSI prediction accuracy as a performance KPI in RAN4.**Proposal 5: RAN4 to consider CSI prediction accuracy metric based on measured CSI in testing of performance monitoring mechanism considering the RAN1 specification.****On alignment of test decoder parameters for Option 3****Observation 5**: In the current RAN4 requirements on CSI/PMI feedback (TS 38.101-4), the tests are separated based on the following parameters: duplexing (FDD/TDD), number of RX&TX (i.e., antenna configuration), type of CSI, and number of PMI.**Proposal 6: RAN4 to use configurations based on one of the existing PMI reporting tests from TS 38.101-4 to evaluate the convergence of AI/ML based reporting performance provided by different companies, e.g., 6.3.3.2.4 Single PMI with 32TX TypeI-SinglePanel Codebook or 6.3.3.2.6 Multiple PMI with 16Tx Enhanced Type II Codebook.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Parameter** | **Description/Examples** | **Nokia preferred parameter values** |
| Model architecture parameters | Model type | Transformer, CNN, RNN, MLP | Transformer |
| Model depth | Number of layers | Several multi-head attention layers (min: [3], max: [7]) |
| Layer type | Fully connected, convolutional, activation layer, etc. | Fully connected layers with activation function for each attention layer/block.*Note that output layer can be different.* |
| Layer size | Neuron count and configuration | Specify embedding and feedforward dimensions, number of attention heads per attention layer/block. |
| Quantization method for the encoder output | Scalar, vector (with codebook) | Scalar quantization  |
| Encoder-decoder interface | Number of latent variables and formatting of bits. | FFS, e.g.,64 latent dimensions with two-bit quantization, i.e., 128 overhead bits. |
| Fixed point representation | Int8, int16, floating point etc | FFS,Decision to be made during/after model design, or may be left for implementation. |
| Format of input to encoder/output of decoder | Eigenvectors, channel matrix, Type II reporting. | Eigen vectors,Sub-band reporting (e.g., [13] sub-bands for 10 MHz CBW, 15kHz SCS.). |
| Model Training related parameters | Training procedure | FFS (e.g Initialization method, training duration, training completion criteria, collaboration type, encoder assumption, etc) | Collaboration type: Type-3 Network first training |
| Loss function | SGCS, NMSE, etc. | SGCS |
| Training datasets | Channel model, number of Tx/Rx portsOther parameters FFS (e.g. rank) | Channel model for training: UMa*Note that in the performance test TDL or CDL (if available) model to be used.*Number of Tx/Rx ports:4 RX,16 or 32 TX,*Note that other options should not be precluded but better to agree on a single scenario as a starting point.*Rank: 1.Channel estimates:Channel eigenvectors derived from [ideal, non-ideal] channel estimates, magnitude normalized to unit length.Dataset size:Sufficient number of samples to achieve minimum performance and prevent underfitting are needed. |
| Hyperparameters | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc. | FFS*since these details depend on selected architecture.* |
| Cross-validation details | Dataset splits for training/testing  | 80%/20%,where training data is also used for validation. |
| Generalization (may be applicable to all ~~four~~ options) | Parameters for Generalization Scenarios | UE speed, SINR, Indoor/outdoor, LOS/NLOS, Propagation model, etc. | FFSdepending on the training assumptions. |
| Performance requirements on test dataset(s) | Mean SGCS, throughput, etc. | FFS, on how to compare performance in identified and other scenarios. |
| Scalability (may be applicable to all ~~four~~ options) | Scalability parameters~~Supported antenna port configurations~~ | Supported antenna port configurations (e.g., (2,8,2), (2,4,2),carrier frequency,bandwidth, etc. | FFS depending on the training assumptions. |
| Supported feedback payloads | Low, medium, high overhead (with specified number of bits) | FFS depending on RAN-1 agreements. |

Table 1: Preferred Parameters Values for Test Decoder Option-3**Proposal 7: RAN4 to consider the preferred values of parameters as described in the Table 1 above for the feasibility checking of Option -3 for AI/ML test decoder design.****On parameters for Option 4:****Observation 6**: Parameters used to specify Option-3 which affect interoperability of the test decoder are likely to affect interoperability if they are used to specify Option-4.**Proposal 8: Parameters specified for Option-3 which affect interoperability, and which are also used to specify Option-4 should use the same values, unless there is clear justification for a change.****Proposal 9: A partially specified dataset is an appropriate choice for defining the training dataset for Test decoder Option-4.****Observation 7**: Eliminating ambiguity in training data samples can be achieved by fully specifying the training dataset. With the implementation of this option, there is no need for performance verification tests. This approach is well-suited for Option-3 of the Test decoder.**Proposal 10: A complete data set dataset is an appropriate choice for defining the training dataset for Test decoder Option-3.****Proposal 11: One of the following approaches can be adopted for freezing a set of parameters for Option-4 from Option-3:*** **Option 1 (Model architecture-based):**
	+ **Option-1a: Freeze a complete model architecture while leaving training data for implementation.**
	+ **Option-1b: Freeze a backbone of model architecture while leaving complete training data and model architectural details for implementation.**
* **Option 2 (Dataset based):**
	+ **Option-2a: Freeze complete training data while leaving model architecture for implementation.**
	+ **Option-2b: Freeze the important characteristics of training data, e.g., number of bits of latent message while leaving actual data samples and model architecture for implementation.**
* **Option 3: Freeze the important characteristics of training data, e.g., number of bits of latent message, and a backbone of model architecture while leaving actual data samples and architectural details for implementation.**

**On encoder and decoder for requirements:****Observation 8**: Even when fixed decoder (like Option-3 of test decoder design) is used for the derivation of requirements a large difference in the performance results might be expected when this decoder is used with company-specific encoders.**Proposal 12: RAN4 to discuss how to ensure alignment of CSI reporting performance and whether to introduce some additional assumptions on the AI/ML encoder-decoder pair parameters for the derivation of the requirements.****On Generalization and Scalability:****Observation 9**: Generalization parameters like the overall scenario (LOS, NLOS, indoor, outdoor, etc.), SINR, UE speed, etc. are generally not known at the UE nor the gNB. And this must be configured at the TE.**Observation 10:** Scalability parameters are generally known at the UE and the gNB and, typically, do not change during the active time of a UE in a certain cell.

|  |  |
| --- | --- |
| **Parameters** | **Description** |
| UE Speed | Slow / Medium / Fast |
| SINR | Good / Bad Radio conditions |
| Outdoor / Indoor | Position of the UE |
| LOS/NLOS |  |
| Propagation Model | UMa / Umi |

Table 2: Parameters for Generalization Scenarios**Proposal 13: For the verification/testing of generalization related aspects in RAN4 for AI/ML enabled CSI feedback enhancement, RAN4 should define different scenarios based on parameters listed in Table 2 above.**

|  |  |
| --- | --- |
| **Parameters** | **Description** |
| Number of Antenna ports | (N1/N2/P) and/or antenna port numbers (e.g., 32 ports, 16 ports) |
| Carrier Frequency | FDD, TDD at sub-band level |
| Bandwidth | E.g., 10MHz, 20MHz |

Table 3: Parameters for Scalability Scenarios**Proposal 14: For the verification/testing of scalability related aspects in RAN4 for AI/ML enabled CSI feedback enhancement, RAN4 should define different scenarios based on parameters listed in Table 3 above.****Proposal 15: As a next step RAN4 should study performance degradation in other scenarios based on simulations and define tolerance margins with respect to identified ones.****On LCM for CSI-feedback:****Observation 11**: 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 16: Core requirements should be considered to limit latency of LCM actions (e.g. activation, deactivation, fallback, switching etc.) typical for the CSI feedback enhancement use case.** |
| [**R4-2405894**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_110bis/Docs/R4-2405894.zip) | Keysight Technologies UK Ltd | ***Proposal 1****: Use relative throughput for CSI prediction inference.****Proposal 2****: Regarding CSI compression sub-use case option 3 (test decoder fully defined in RAN4 specifications) agree that “Encoder-decoder interface”, “fixed point representation”, “format of input to encoder/output of decoder” and model training procedure parameters “training completion criteria”, “collaboration type” and “encoder assumptions” are required to check the feasibility of this option.****Proposal 3****: Regarding CSI compression sub-use case option 4 (test decoder partially defined in RAN4 specifications and TE vendors in charge of final implementation of test decoder), RAN4 to restrict the discussions to current “Option 1 (4a-1) + Option 3” or “Option 2 + Option 3” or just focus on the latter one.****Proposal 4:*** *Regarding CSI compression sub-use case option 4 (test decoder partially defined in RAN4 specifications and TE vendors in charge of final implementation of test decoder), RAN4 to discuss definition of parameters required to check the feasibility of this option, including the evaluation of equivalent test decoder implementation by different TE vendors.* |

## Open issues summary

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

1. CSI prediction KPIs
2. Reference decoder and test decoder
3. Option 3 split
4. RAN4 – RAN 1 alignment
5. standardization steps for Option 3
6. Option 3 for 2-sided model – parameter discussion
7. Option 4 for 2 sided model
8. TE verification for option 4

### Sub-topic 4-1

*CSI prediction KPI*

Some companies brought somewhat different options for CSI prediction test metric.

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

* Proposals
	+ Option 1: Use relative throughput (e.g. predicted PMI vs random PMI) as the test metric
		- use Type I single-panel,
	+ Option 2: Use absolute minimum throughput as metric
	+ Option 3: Others
* Recommended WF

To be discussed

### Sub-topic 4-2

*Reference encoder/decoder and Test encoder/decoder*

There is some confusion on the terminology of reference encoder/decoder and test encoder/decoder, this should be clarified

**Issue 4-2: Reference and test encoder/decoder**

* Proposals
	+ Option 1:
		- Reference encoder/decoder: the encoder/decoder used in RAN4 discussions for simulation alignment/requirement derivation. It could be documented (in TR, WF, etc) or captured in the specifications as necessary.
		- Test encoder/decoder: the encoder/decoder to be used in RAN4 tests. it will be captured in the specifications if necessary(for example, for Option 3 it would be captured in the specifications)
	+ Option 2: Other definitions
* Recommended WF
	+ Option 1

To be discussed any other changes are needed

### Sub-topic 4-3

*Option 3 split*

The possibility of splitting option 3 was brought up

**Issue 4-3: Option 3 split**

* Proposals for parameters that should be agreed upon:
	+ Option 1:
* Option 3a: The test decoder is determined for each test case, using a specific dataset collecting from the configuration/scenario of the considered test case.
* Option 3b: The test decoder is determined for more than one test cases, using a mixed dataset collecting from different configurations/scenarios of the considered test cases.
	+ Option 2: Decision to split should be made later (in an eventual WI phase) based on feasibility
		- it will depend on whether it is feasible to have a “one fits all” decoder or whether it has to be scenario specific
	+ Option 3: others
* Recommended WF
	+ Option 2

### Sub-topic 4-4

*RAN4 – RAN1 Coordination*

Some companies brought up the fact that the RAN1 discussion on interoperability for the 2 sided model is similar to the RAN4 discussion on Options 3/4 and some further alignment would be needed

 **Issue 4-4: RAN4 – RAN1 Coordination**

* Proposals
	+ Option 1: Further alignment is needed between RAN4 and RAN1
		- which parameters to align?
	+ Option 2: Discussion in RAN4 should continue, RAN4 should determine testing feasibility of Option 3/4
	+ Option 3: Option 1 and Option 2
	+ Option 4: others
* Recommended WF
	+ To be discussed

Concrete proposals on which parameters to align would be useful

### Sub-topic 4-5

*standardization steps for Option 3*

Further steps on how Option 3 would be standardized should be discussed. This discussion should apply to Option 4, at least partially. A flow chart is provided in R4-2405653 and reproduced below for convenience:

**Step-1: Identify necessary Model Architecture Parameters**

**Standardization Procedure End**

for a certain use case
(e.g., CSI compression for precoding matrix under certain config.)

Model architecture parameters could include: Model type, Model depth, Layer type/size, Quantization, etc.

Model training procedure, loss function, training datasets, hyperparameters, etc.

**Step-3: Companies provide two-sided model design based on their own study/preference**

**Step-4: Performance comparison based on different companies’ en/decoder designs**

**Yes**

**No**

**Step-6: Performance alignment by companies based on agreed model architecture/training parameters**

Performance comparison in terms of metrics like NMSE, SGCS, etc.

**No**

**Yes**

**No**

**Standardization
Procedure Start**

**Step-2: Identify necessary Model training Parameters**

Test decoder is expected to be generated in this step

Reference encoder is assumed, but leave enough implementation flexibility to vendors (similar to Demod alignment for MMSE-IRC)

**Step-8: performance alignment
for encoder design by companies
based on assumptions on reference encoder**

**Yes**

RAN4 performance requirement obtained (for certain reference encoder)

**Step-10: Derive RAN4 performance requirement**

**Step-5: RAN4
agree on two-sided model architecture
 / training parameters?**

**Step-7: RAN4 agree on test decoder
(which can be fully specified in spec.)**

**Step-9: RAN4 achieve performance alignment?**

**Issue 4-5: Option 3 standardization process**

* Proposals
	+ Option 1: This flowchart can be used as reference, RAN4 is now discussing Steps 1-4, RAN4 to futher work on feasibility to pass Step - 5 and move to Step - 6
	+ Option 2: others
* Recommended WF

To be discussed:

Current status and further steps to align the understanding and goals among companies

### Sub-topic 4-6

*Option 3 for 2-sided model*

Discussion on choosing parameters for the option 3 feasibility study started in the previous meeting and some agreements were captured in a table reproduced below (agreements highlighted in green, tentative agreements highlighted in yellow):

|  |  |  |
| --- | --- | --- |
| **Category** | **Parameter** | **Description/Examples** |
| Model architecture parametersa | Model type | Transformer, CNN, RNN, MLP |
| Model depth | Number of layers |
| Layer type | Fully connected, convolutional, activation layer, etc. |
| Layer size | Neuron count and configuration |
| Quantization method for the encoder output | Scalar, vector (with codebook) |
| Encoder-decoder interface | Number of bits of latent message |
| Fixed point representation | Int8, int16, floating point etc |
|  | Format of input to encoder/output of decoder |  |
| Model Training related parameters | Training procedure | FFS (e.g Initialization method, training duration, training completion criteria, collaboration type, encoder assumption, etc) |
| Loss function | SGCS, NMSE, etc. |
| Training datasets | Channel model, number of Tx/Rx portsOther parameters FFS (e.g. rank) |
| Hyperparameters | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc. |
| Cross-validation details | Dataset splits for training/testing/validation |
| Generalization (may be applicable to all four options) | Performance requirements on test dataset(s) | Mean SGCS, etc. |
| Scalability (may be applicable to all four options) | Supported antenna port configurations | (2,8,2), (2,4,2), etc. |
| Supported feedback payloads | Low, medium, high overhead (with specified number of bits) |

Proposals for the other parameters were brought in this meeting and these need to be discussed.

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

* Proposals for parameters that should be agreed upon:
	+ Option 1: The proposals from different companies are captured in the table below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Proponent | Vivo | CATT | SEU | QC | Xiaomi | Intel | E/// | Apple | Nokia | Keysight |
| Model type | Transformer and CNN | Transformer | CsiNet | MLP | Transformer | Several model types can be considered (e.g., transformer, CNN)RAN1 inputs on the best identified models in terms of performance/complexity can be requested  | Transformer or CNN depending on design target\*\*In order to decide on the model structure, it is necessary to consider RAN1 evaluations and also consider whether the RAN4 standardized test decoder should target maximum performance or some complexity limit.RAN4 should decide and agree on whether the target is purely performance or both performance and complexity. (Some checking with RAN1 may be needed). | CNN | Transformer | 　 |
| Model depth | 　 | 　 | 　 | Three linear layers (with one activation function) | 6 | decide upper bound of complexity based on RAN1 evaluation | 　 | Number of layers, CNN: Kernel/Filter Size, Padding, Stride, Pooling layers parameters, Number of channels | Several multi-head attention layers (min: [3], max: [7]) | 　 |
| Layer type | 　 | 　 | 　 | MLP with expansion factor N = 4, and each layer/function is described in the following• 1st linear layer: input is latent message of size Zdim and output is a vector of size nSB x nTx• Reshape: convert the vector of size nSB x nTx to nSB vectors with size nTx• 2nd Linear layer: For each subband, the input is a vector of size nTx, and the output is a vector of size N x nTx. The same linear layer is applied to each of nSB subbands.• Activation function: GELU• 3rd Linear layer: the input is a vector of size N x nTx, and the output is with a vector of size nTx. The same linear layer is applied to each of nSB subbands. | Scalar | 　 | 　 | Fully connected, convolutional, activation layer (activation type: leakyRelu,etc), batch(group)-normalization layer,dropout layer, etc. | Fully connected layers with activation function for each attention layer/block. | 　 |
| Layer size | 　 | 　 | 　 | 　 | 60, 120, 280bits, 240 bits for generalization | 　 | 　 | Neuron count and configuration | *Note that output layer can be different.* | 　 |
| Quantization method for the encoder output | 　 | 　 | 　 | scalar quantizer, 2 bits per dimension (element) | int16 | 　 | 　 | Scalar, vector (with codebook) | Specify embedding and feedforward dimensions, number of attention heads per attention layer/block. | 　 |
| Encoder-decoder interface | 　 | 　 | 　 | 　 | 60, 120, 280bits | 　 | Consider 63, 110 or 230 | Number of bits of latent message | Scalar quantization  | 　 |
| Fixed point representation | 　 | 　 | 　 | 　 | 　 | 　 | Int8, int16, floating point etc. | Int8, int16, floating point etc | 　 | 　 |
| Format of input to encoder/output of decoder | 　 | 　 | 　 | 　 | 　 | 　 | Consider pre-processing of Eigenvector using Enhanced Type 2 codebook | 　 | 　 | 　 |
| 　 | 　 | 　 | 　 | 　 | 　 | 　 | 　 | 　 | 　 | 　 |
| Training procedure | 　 | 　 | 　 | 　 | 　 | 　 | Convolutional: Feedback bits per transmission e.g., 10 x 4 = 40Transformer: Optimizer, e.g., Adam | collaboration training type need to be specified | Collaboration type: Type-3 Network first training | Training completion criteria is probably one of the most important parameters assuming it will include boundaries (minimum and maximum) for the test decoder performance required.Collaboration type will determine interactions required between different stakeholders and/or different AI/ML algorithms blocks |
| Loss function | 　 | 　 | 　 | SGCS | 　 | 　 | NMSE | 　 | 　 | 　 |
| Training datasets | 　 | 　 | 　 | Encoder input dataset should cover all the contributing companies’ encoder input data | 　 | 　 | 　 | Number of layers/rank?SNR, Genie/ real channel estimates (impairments)?Data format of training (depends onCollaboration training type) Size of training data setSpecify channel model parameters or training data samples stored in a repository?Different Training Sets (configurations/ scenarios)?Multiple vendor training sets  | Channel model for training: UMaNote that in the performance test TDL or CDL (if available) model to be used.Number of Tx/Rx ports:4 RX,16 or 32 TX,Note that other options should not be precluded but better to agree on a single scenario as a starting point.Rank: 1.Channel estimates:Channel eigenvectors derived from [ideal, non-ideal] channel estimates, magnitude normalized to unit length.Dataset size:Sufficient number of samples to achieve minimum performance and prevent underfitting are needed. | 　 |
| Hyperparameters | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc.Note that training procedure does not need to be fully aligned. | 　 | 　 | 　 | Learning rate = 0.001,batch size = 128optimization algorithm = Adam,  | 　 | 　 | 　 | 　 | 　 |
| Cross-validation details | Dataset splits for training/testing/validation | 　 | 　 | 　 | Dataset for training: 199,500Dataset for Testing: 10000Dataset for validation: 10500 | 　 | 　 | Dataset splits for training/validation/testingThis testing doesn’t refer to DUT testing | 80%/20%,where training data is also used for validation. | 　 |

* Recommended WF
	+ To be discussed

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

### Sub-topic 4-7

*Option 4 for 2-sided model*

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

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

* Proposals
	+ Option 1: Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder
	+ Option 2: discuss which parameters/assumptions should be different/not needed compared to Option 3 in Issue 4-6, one example below (Nokia – R4-2405738)
* **Option 1 (Model architecture-based):**
	+ **Option-1a: Freeze a complete model architecture while leaving training data for implementation.**
	+ **Option-1b: Freeze a backbone of model architecture while leaving complete training data and model architectural details for implementation.**
* **Option 2 (Dataset based):**
	+ **Option-2a: Freeze complete training data while leaving model architecture for implementation.**
	+ **Option-2b: Freeze the important characteristics of training data, e.g., number of bits of latent message while leaving actual data samples and model architecture for implementation.**
* **Option 3: Freeze the important characteristics of training data, e.g., number of bits of latent message, and a backbone of model architecture while leaving actual data samples and architectural details for implementation.**
* Option 3: standardize a reference encoder and a test data framework and test dataset/channel model (Ericsson – R4-2405611))
	+ Option 4 – Huawei (R4-2405151)
* Option 4a: Model structure is not specified in RAN4. Training dataset is specified, where each training sample consists of both the raw channel matric/precoding matrix and the bit stream forwarded to the test decoder.
* Option 4b: Model structure is specified in RAN4. Training dataset is not specified for verifying the encoder at DUT. The test decoder developed by TE vendor needs verification.
	+ FFS: How to determine the test metric for test decoder developed by each TE vendor.
	+ Option 5: Qualcomm (R4-2404216)
* **Option 4a-1 standardized dataset**
	+ **Step 1: RAN4 agrees a pair of encoder and decoder with full details (same as fully specified decoder discussion) and an encoder input data generation procedure.**
	+ **Step 2: RAN4 uses this encoder/decoder pair and the generation procedure to generate a set of decoder input and output data and captures this dataset in the specification.**
	+ **Step 3: RAN4 specifies a test decoder verification procedure based on the specified dataset.**
* **Option 4a-2 standardized aggregated dataset**
	+ **Step 1: RAN4 achieves some agreements (e.g., part of but not all the parameters in the test decoder parameter table in the previous meeting WF[1]) for the test decoder.**
	+ **Step 2: Interested companies can design their own encoder/decoder pairs based on the agreements to contribute the (decoder input, decoder output) dataset to RAN4**
	+ **Step 3: RAN4 aggregates the datasets from all the contributing companies, and capture the aggregated dataset in the specification**
	+ **Step 4: RAN4 specifies a test decoder verification procedure based on the specified dataset.**
* **Option 4b reference encoder**
	+ **Step 1: RAN4 agrees a pair of encoder and decoder with full details (same as fully specified decoder discussion) and an encoder input data generation procedure**
	+ **Step 2: RAN4 capture the encoder as a reference encoder and the encoder input data generation procedure in the specification.**
	+ **Step 3: RAN4 specifies a test decoder verification procedure based on the reference encoder.**
	+ Option 6: Others
* Recommended WF
	+ To be discussed

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

### Sub-topic 4-8

*TE verification/validation for Option 4*

Option will leave the actual test decoder implementation to the TE vendors, there might be a need to verify that the decoder is correctly implemented and has the right level of performance.

**Issue 4-8: TE decoder verification/validation**

* Proposals
	+ Option 1: RAN4 will have to come up with a TE verification/validation procedure
		- details are FFS
	+ Option 2: This should be left to RAN5
	+ Option 3: TE verification/validation is not needed
	+ Option 4: other options
* Recommended WF

To be discussed