**3GPP TSG-RAN WG4 Meeting #111 R4-2408944**

**Fukuoka city, Japan , May 20th – 24th, 2024**

**Agenda item:** 10.11.5

**Source:** Moderator (Qualcomm)

**Title:** Topic summary for [111][133] NR\_AIML\_air

**Document for:** Information

# Introduction

This is the summary thread for issues related to NR AI/ML study in RAN4. The Rel-19 WID was agreed in RP-234039. This is a continuation of the Rel-19 SI. A WF listing some agreements and several issues for further study was agreed in R4-2406617 in RAN4#111.

# 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** |
| [**R4-2407233**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407233.zip) | Apple | **Observation 1: "Model as a baseline" can either be explicitly captured in RAN4 specifications or agreed upon for aligning performance results.****Observation 2: For Option 1 and for verifying DUT’s AI/ML capability to load and execute the models, trained AI/ML models can be defined in RAN4 spec for different use case tests****Observation 3: For Option 1 and for ensuring the model is properly conducted, performance requirements can be established for RAN4-defined AI/ML models across various use cases. The DUT is considered to have successfully passed the tests if it meets the specified performance requirements.****Observation 4:** **The identified scenarios and configurations can be initially understood as those reported by UE through capability signaling** **as part of functionality identification.****Observation 5: Different scenarios that will be part of generalization test could act as the additional conditions for the AI/ML model training but do not constitute a part of UE capability for the AI/ML-enabled feature/FG****Observation 6: Configurations utilized for generalization test should be associated with UE capability of an AI/ML-enabled Feature/FG (set the configuration and vary the conditions under the configuration). Different conditions should be applied for each configuration.****Observation 7: The existence of a wide range of diverse UE capabilities poses a challenge for RAN4 in identifying a typical configuration or scenario for specifying test cases.****Observation 8: On device fine-tuning based on retraining could be beneficial to model delivery/transfer to reduce overhead/latency****Observation 9:** **Having a separate AI/ML model for each unique scenario, configuration, and additional condition could significantly increase the complexity and storage requirements of the User Equipment (UE). Additionally, it may introduce overhead in terms of model delivery, transfers, and associated latency****Observation 10: The post-deployment phase can be considered within the broader context of the generalization framework****Observation 11: Options 1-a and 1-b for post deployment testing would impose a large burden for testing, signal overhead and complexity.*****2.1 RAN4 performance testing goals*****Proposal 1: Deprioritize Option 1 and focus on Option 2 for performance requirements as a testing goal** **Proposal 2: For verifying performance gain of AI/ML models/functionalities, RAN4 can define multiple independent** **test cases with different scenarios and configurations/conditions as reported through UE capability signaling, which could include:** **Propagation conditions/scenarios, e.g., channel modes defined for different scenarios (CDL, AWGN, etc.) in TR38.901, Doppler conditions, SNR levels, various deployment scenarios (Uma, Umi, ISD), different cell with different gNB height, various indoor/outdoor UE distributions, various UE mobility etc.** **- Configurations:*** **Various UE parameters: number of UE beams, antenna panels, antenna array dimensions, different AI/ML configurations: different set B of beams, T1 for measurement/T2 for prediction for BM-case2**
* **Various gNB settings: Tx codebooks (set A/B beams), beam widths, antenna spacings, Tx antenna dimension, antenna port layouts, TXRU antenna virtualization (e.g. (8,8,2,1,1,2,8), vs (8,4,2,1,1,2,4))**

**RAN4 should conduct a down-selection process to narrow down the scenarios and configurations (parameters) for the conformance test. This test will serve as a benchmark for evaluating the model's performance across a range of unseen scenarios and configurations.** ***2.2 Static/non-static scenarios/configurations*****Proposal 3: If non-static scenarios/configurations are supported for certain use cases, they can be included as part of generalization tests.****Proposal 4:** **Non-static scenarios/configurations should be considered for test cases only if static scenario/configuration testing fails to fulfill the testing objectives. CSI and beam management temporal prediction use cases are particularly suitable for introducing non-static environments during testing****2.3 Generalization/scalability aspects****Proposal 5: To determine the granularity of additional scenarios/conditions for defining generalization tests for each use case, it's essential to study how the AI model's behavior changes with different scenarios and conditions.** **Proposal 6: To improve the generalization behavior of the model, training with a diverse dataset should be investigated.** **Proposal 7: Investigate the reduction of generalization tests by training with a mixed dataset containing samples from different configurations. Investigate the definition of a single generalization test where the testing data consist of a random mixture of these configurations. If the performance degradation between the model trained on the mixed dataset and tested with random mixture configurations and the model trained and tested specifically for each configuration meets a predefined performance margin criterion, the test could pass.**Proposal 8: Study post-deployment procedures to augment conformance testing for effectively managing performance across all possible deployment conditions/scenarios (which are not tested)**Proposal 9: For UE-side models and/or UE-part of two-sided models** **it is suggested that the scenarios and configurations utilized for generalization tests can be determined based on the supported configuration reported by the UE as part of capability signaling.****Proposal 10: In the RAN4 core requirement, it is mandated that the consistency or association between of additional conditions during both training and inference is guaranteed. This serves as additional assistance information from the network side for testing UE-sided models during inference and monitoring****Proposal 11: RAN4 will explore methods to specify the identified scenarios and/or configurations per use case in future release, contingent upon other WGs can specify the granularity and the capability signaling.****Proposal 12: For defining generalization tests, RAN4 should define identified scenarios associated with the UE capability report of an AI/ML-enabled Feature FG, and other scenarios (additional conditions). RAN4 should also define minimum level of performance for the identified scenarios and/or conditions.** **Proposal 13: Other scenarios and/or configurations can be interpreted as the scenarios and/or configurations that are not reported by UE capability signaling for an AI/ML-specific functionality or model ID.***2.3.1 Scenario/configuration specific Models (Fine-tuning)***Proposal 14: RAN4 to discuss the practicality of formulating a framework that facilitates on-device fine-tuning. The focus will be on exploring the feasibility of creating a dynamic and site-specific approach to online training and fine-tuning (e.g reinforcement learning)** **Proposal 15: UE to update its stored AI/ML models with the new model ID after fine-tuning, where this ID can be associated with the training data (which implicitly have the additional conditions) used to fine tune the model.** Proposal 16: RAN4 should investigate the options for enhancing the generalizability of AI/ML models by providing the appropriate assistance/side information as input signal to the inference engine of the AI/ML model and discuss the feasibility of training with diverse datasets across different additional conditions*2.4 Principles on the definition of requirements*Proposal 17: RAN4 should clarify/agree that the side conditions of the testing procedures should remain the same for legacy and AI/ML methods.*2.5 Requirements for LCM (Performance Monitoring)***Proposal 18: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case** **Proposal 19: RAN4 shall consider the latency requirements for model monitoring input data as well as the establishment of tolerance margin requirements for the specified KPIs for model monitoring per use case****Proposal 20: FFS on how to perform cell level BM performance monitoring when the AI/ML model resides at NW***2.6 Post Deployment validation and Fine tuning* **Proposal 21: If the AI/ML model drifts due to misalignment of network-side additional conditions, the alignment could be achieved through an ID assigned by the network during training data collection. This ID indicates the association of the training data with the additional conditions implied to generate those data** **Proposal 22: If the AI/ML model drifts due to misalignment of network-side additional conditions, if the UE lacks the ID but the network possesses the model ID, then the network can transfer the model to the UE, and the UE can update its list of models****Proposal 23: If the AI/ML model drifts due to misalignment of network-side additional conditions, if the UE has a model that partially supports the network-sided additional conditions, then the network can trigger data collection for fine-tuning the UE-sided model. Subsequently UE updates its list of model IDs.****Proposal 24: The UE updates its model list with a new model derived from either fine-tuning, model transfer from the NW, or monitoring to ensure the consistency of additional conditions. Then, the UE assigns an ID to the new model that supports the NW's additional conditions and shares this information with the NW. If some of the new conditions/configurations are standardized, the UE updates its capability report accordingly.****Proposal 25: If there is no ID available to associate training data with additional conditions, and monitoring procedure fails to guarantee the consistency of the model with the additional conditions then UE should fallback to legacy mode.** **Proposal 26: Study an on-device model monitoring/performance assessment to monitor active and inactive AI/ML models with the actual UE hardware and field data to dynamically manage the database of models and enable proactive monitoring that facilitates seamless transitions to newly updated or introduced models when needed, ensuring efficient model management.***2.7 Data collection for testing***Proposal 27: 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.** |
| [**R4-2407366**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407366.zip) | CAICT | **Proposal 1: Consider effective performance monitoring for post-deployment handling of minor model update or fine-tuning.****Proposal 2: Whether and how to define test on model transfer/update shall wait for more explicit progress from other WGs.****Proposal 3: Suggest to discuss potential solutions for offline performance tests conducted inside vendor.****Proposal 4: Some initial views on static/non-static conditions for different use cases and tests are presented in table 1.****Table 1 Initial views on static/non-static conditions for different use cases and tests**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use case** | **Performance test** | **LCM test** | **Generalization test** |
| **CSI compression** | * Static in terms of test parameters as legacy PMI reporting test
* For CSI prediction, a larger maximum doppler frequency and LOS path could be considered in propagation condition
 | Three aspects in general:* Monitoring test: the main body is measurement and reporting
* Decision-making: up to vendor implementation (may not be a part of RAN4 test)
* Management: the main body is signaling procedure

**Observation:**Following the understanding above, LCM test seems to be irrelevant to static/non-static conditions. But, non-static conditions that would induce significant performance fluctuation could be introduced to trigger model switch/fallback/deactivation. | * The test procedure is same as that of performance test
* Be non-static in terms of some key aspects (generalization evaluations in TR38.843 can be starting point)
* How to introduce non-static conditions in the test is FFS:
	+ Alt 1: randomly change (feasibility on TE side needs clarification)
	+ Alt 2: sequentially change in predefined order (difference compared to splitting into individual static test requires clarification)
 |
| **CSI prediction** |
| **BM-case 1** | * Static in terms of test parameters as legacy L1-RSRP measurement test other than propagation conditions
* Non-static for propagation conditions to emulate different beams and UE movement
 |
| **BM-case 2** |
| **AI-based positioning** | Waiting for more RAN1/RAN2 progress |

 |
| [**R4-2407376**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407376.zip) | NTT DOCOMO, INC. | **Proposal 1: Static and non-static scenario/configurations are summarized in table below.**

|  |  |  |  |
| --- | --- | --- | --- |
| Use case | Feature | Static/non-static | Non-static parameters |
| Beam management | Spatial-domain downlink beam prediction (BM-Case 1) | Static | Specific channel realizations also static |
| Temporal downlink beam prediction (BM-Case 2) | Non-static | SNR (varied according to trajectory), [Doppler shift, LOS/NLOS] |
| Positioning | Direct AI/ML positioning | Static | Specific channel realizations also static |
| AI/ML assisted positioning | Static | Specific channel realizations also static |
| CSI prediction | Spatial-frequency domain CSI compression | Static | Specific channel realizations also static |
| Time domain CSI prediction | Non-static | SNR (varied), [LOS/NLOS] |

 |
| [**R4-2407496**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407496.zip) | CATT | **Proposal 1: Companies clarify the scope of changes/scenarios that need to be verified by post deployment tests.** **Observation 1: Regarding the framework in option 2, the following two aspects are difficult to be verified:** **- Whether device vendors have performed tests on changes to UE devices and/or,****- Whether self-test results conveyed to network by device vendors are reliable.** **Proposal 2: The performance of AI/ML models after changes are implemented is ensured by device vendors and the performance monitoring procedure.** **Proposal 3: Use static scenarios/configurations as baseline for tests.** **Proposal 4: Refine the definitions of static and non-static scenarios/configurations during the design of test cases.** **Proposal 5: Postpone the discussion on inference latency requirements until more details are defined by RAN1/2.** |
| [**R4-2407845**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407845.zip) | Xiaomi | **Observation: It’s FFS whether model monitoring can be used for model performance test. There are several possible options for performance monitoring. RAN4 needs more input from other WGs.** **Proposal: It’s FFS whether model monitoring can be used for post-deployment performance test. RAN4 needs more input from other WGs.**  |
| [**R4-2408177**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408177.zip) | CMCC | ***Proposal 1: it is proposed to firstly focus on static scenarios/configuration considering the timeline and workload. Non-static scenarios/configurations can be further discussed if time allowed.*** ***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.***  |
| [**R4-2408291**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408291.zip) | vivo | **Observation 1: At least for BM case 2, use non-static scenarios/configurations as baseline.****Proposal 1: Consider the following aspects as the factors for non-static test:*** **Scenarios condition, e.g., UMa, UMi-Street Canyon, RMa, InH-Office or InF**
* **SNR**
* **Channel condition, e.g., LOS/NLOS, and small scale parameter, e.g., the cluster power, the cluster number and the cluster arrival/departure angles.**
 |
| [**R4-2408491**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408491.zip) | Ericsson | Proposal 1 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 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. (Examples of the common set of procedures include: identification, selection, activation, deactivation, switching, and fallback to non-AI operation, as well as performance monitoring.)Proposal 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.Proposal 4 At least the legacy RRM requirements (non-AI/ML) have to be met, even during the AI/ML operation mode.Proposal 5 A maximum delay for switching from non-AI/ML and back to non-AI/ML operation can be a candidate RRM requirement.Proposal 6 RAN4 further discuss the whether model merging is feasible and useful or not as a possibility for agreeing a reference or test model. |
| [**R4-2408615**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408615.zip) | Intel Corporation | **Proposal #1: UEs supporting AI/ML features shall be mandated to meet at least the subset of the existing legacy performance requirements with configured/enabled AI/ML functionality.****Proposal #2: Consider the following framework for post-deployment model (feature/functionality) verification:*** + **UE shall support at least some default/baseline AI/ML models/features that passed conformance requirements. The respective models can be used as a fallback mechanism.**
	+ **Changes or updates to the AI/ML models/features may be tested by the device vendors against RAN4 requirements before the deployment to the UE is performed. The information on whether AI/ML model/feature update has passed conformance test (and potentially associated data) shall be conveyed to the network, and based on this, the network may adjust the model monitoring framework accordingly.**
	+ **Further discuss a potential specific performance validation procedure during AI/ML model transfer/update as a part of performance monitoring framework (involving RAN1).**

**Proposal #3: Define reference AI/ML models for one-sided and two-sided models (both encoder/decoder) for performance requirements definition taking into account performance/complexity tradeoffs.**  |
| [**R4-2408658**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408658.zip) | Nokia | ***On testing in non-static scenarios/configurations:*****Observation 1:** Non-static scenarios in RAN4 are required to verify that there is no performance degradation while UE may autonomously switch to different AI/ML models. Different use cases may have different requirements and tests for non-static scenarios due to different performance impacts during transition period of model switching.**Proposal 1: RAN4 should define performance requirements for non-static scenarios/conditions provided that the functionality is not changing/switching during the test and internal model’s transitions are up to the UE.*****On post-deployment handling:*****Observation 2:** Frequent updates to the models/functionality may significantly prolong the turnaround time and testing overhead.**Proposal 2: Conformance testing for updated AI model/functionality before deployment (Option 1) is feasible in the case of infrequent declared centralized changes/updates to the models/functionalities (e.g., based on offline (re)training).****Observation 3:** Current Option 2 does not completely address the issue of the use of the models/functionalities in the live network, that have not passed conformance testing.**Proposal 3: RAN4 to ensure (and to clarify in Option2: performance monitoring-based post-deployment handling) that at least one version of the model/functionality that passed conformance testing shall always be available in the device.****Observation 4:** In order to validate the new model/functionality before using it in the filed, the UE need to support inference of several models either in consecutive/time splitting manner or in parallel.**Proposal 4: RAN4 to consider handling post-deployment verification based on monitoring of updated/new functionality/model in inactive/standby mode along with the inference of the currently used model/functionality.*****On data collection requirements:*****Proposal 5: In RAN4, requirements on data collection for training, data collection for inference, and monitoring data collection shall be discussed separately.****Observation 5:** 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 6:** 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 7:** 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 6: RAN4 to focus on the (inference) latency core requirements (i.e., in between the measurements/signalling and reporting) instead of ‘data collection for inference’ requirement and continue the related discussions for each use case separately.****Proposal 7: RAN4 needs to discuss accuracy and latency requirement on data collection for training only if it is triggered by other WGs.****Proposal 8: RAN4 to consider monitoring data collection requirements in a use-case specific manner and based on RAN1 design of the corresponding mechanisms.*****On testing goals and definitions:*****Observation 8:** Testing goal Option 1 is only possible with model identification, and mostly relevant for the specific use-case of two-sided CSI compression. Testing of performance requirement for a known model becomes a formality that only tests the UE hardware.**Proposal 9: Testing according to Option 1 cannot be sufficient to conclude about the conformance of the whole functionality.****Proposal 10: RAN4 to consider the following update for testability goal Option1 if it needs to be kept:Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible), can fulfil minimum requirements specified in RAN4.****Observation 9:** 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 11: 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.*****On reference block diagram for testing functionalities:*****Proposal 12: RAN4 to agree that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 is applicable only for the testing UE-sided use cases,** **and its description requires further clarifications for the definition of the depicted blocks and links/arrows between them as show in the figure below.** |
| [**R4-2409000**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409000.zip) | Huawei,HiSilicon | ***Proposal 1***: Non-static scenarios/configurations is precluded for defining RAN4 test.***Proposal 2:*** 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

***Proposal 3:*** Take functionality-based LCM as the starting point for RAN4 discussion.***Observation 1:*** Identified scenarios and/or configurations can be initially interpreted as the scenarios and/or configurations that UE reports by capability signaling.***Observation 2:*** 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 3:*** If legacy test metrics are not valid/testable when defining AI/ML-specific requirements, legacy performance requirements for non-AI cannot be reused.***Proposal 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 4:*** 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 5:*** There is no benefit identified by introducing channel condition changes during test.***Proposal 9:*** Non-static scenarios/conditions and propagation conditions are precluded for defining RAN4 test.  |
| [**R4-2409464**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409464.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)

*Static/non-static scenarios/conditions for testing***Proposal 2: 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 1: 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 3: Option 1 of post-deployment handling is preferred only if** **- The conformance testing includes the validation testing performed by UE vendors in the development phase of AI/ML model, rather than formal conformance tests in certificated labs.** **Proposal 4: RAN4 should continue the post deployment discussion based on the monitoring progress in other WG while keeping the previous agreement as Option 1 and Option 2.** |
| [**R4-2409686**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409686.zip) | ZTECorporation,Sanechips | **Proposal 1: The scenario of UE capability report, which is AI authentication, it can be identified. Otherwise, the remaining scenarios can be regarded as the other scenarios for test.****Proposal 2: The minimum level of performance shall be studied per use case.****Proposal 3: The legacy performance under different use cases can be as the baseline to judge the minimum performance of AI-based method for each identified scenario.****Proposal 4: The margin shall be added based on the known performance gain for identified scenario to judge what is the significant degradation for other scenarios and try to guarantee that it will not occur**.**Observation 1: 1. The different use case has the different synthetic channels. 2. One synthetic channels shall be used for the individual use case****Proposal 5: RAN4 shall confirm that the different use case has the different synthetic channels.****Proposal 6: RAN shall consider how to design the synthetic testing data.** |

## Open issues summary

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

1. Post deployment testing
2. Static and non-static conditions in tests
3. Static and non-static conditions clarifications
4. Inference Latency handling
5. LCM Requirements
6. Legacy RRM Requirements handling
7. Testing goals update
8. Testing diagram update/clarification
9. Channel models to be used for training/test data

### Sub-topic 1-1

*Post Deployment testing*

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

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

**Agreement:**

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

There was further discussion in RAN4#110Bis but there were no agreements. Several companies brought proposals in this meeting.

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

* Proposals
	+ Option 1: Option 1 is feasible only if model updates are infrequent
	+ Option 2: Option 2 is feasible only if the validation can be performed by UE vendors individually (no need to test in a certified lab)
		- FFS whether some validation information needs to be shared with gNB
	+ Option 3: Add an option in which the conformance test inputs are stored and used for post-deployment testing/validation
	+ Option 4: Drop Option 1, only discuss post deployment handling as part of LCM
	+ Option 5: RAN4 should discuss a framework to ensure that at least a validated model exists at the UE
	+ Option 6: RAN4 should discuss a framework to ensure that non-validated/non-tested models are not used by the UE
	+ Option 7: Others
* Recommended WF

To be discussed, multiple options ca be adopted

NTT DOCOMO: Support option 4 and option 6. “if model updates are infrequent” need clarifications.

Huawei: before talking about the post deployment testing, how can network identify the post deployment procedure and whether the model is post-deployed or pre-deployed? RAN4 cannot collect field data for model update. It is not like the simulation. We prefer option 4.

Ericsson: in our view, this is different issue to be solved. There is no point to do any RAN4 work unless … We proposed to previous option 1 to be clarified further, and new option 3 here.

Nokia: Proposal 6 comes from US and important. How to monitor the model depends on the procedure. Infrequency changes in option 1 is not feasible. We also support option 5.

Apple: It is very important topic. We do not have to prioritize the option at this moment. Can we deprioritize the topic unless we understand the regular test framework?

Samsung: Based on the previous agreement on option 1 and option 2, we are fine the keep both options. We support option 2 and 4.

Xiaomi: We share the same understanding as Apple. It is still FFS about the test metric.

Intel: This is important issue to be addressed. We propose to down-select the options. It is not good approach to push it to very end.

CATT: Support to deprioritize the topic until the LCM is concluded.

ZTE: share the same view as Xiaomi. Monitoring procedure is on-going. We need discuss it per use case.

Ericsson: Encourage companies to continue thinking about it.

Intel: Ericsson suggestion makes sense. LCM may not be sufficient. Otherwise why do we need the performance requirements.

Nokia: all the companies agree the importance. LCM is still discussed under other WGs. We need verify the goal.

Apple: To Intel, this is exact the reason that we think it is not right time to discuss. As long as UE guarantee the performance. We should have clear definition of LCM and what the performance is.

Moderator: suggest to look at whether options can work and is feasible.

### Sub-topic 1-2

*Static and non-static conditions*

In the previous meeting it was agreed that static and non-static scenarios/configuration could be needed for AI testing, however, no further details were agreed.

Agreement:

* Both static and non-static scenarios/configurations could be needed for AI testing
	+ RAN4 will further discuss how to use them case by case
		- FFS whether to use static scenarios/configurations as baseline.
	+ Refine the definitions of static and non-static scenarios/configurations based on two bullets below
		- Static: channel model and SNR settings are fixed and do not change over the test, specific channel realizations may be dynamic
		- Non-static: Non-static scenarios/configuration can be further considered in application to use cases. The details of models are FFS and may include non-stationary SNR and other conditions.

**Issue 1-2: Static and non-static conditions**

* Proposals
	+ Option 1: Non-static scenarios/configuration should not be used for AI/ML testing in Rel-19
	+ Option 2: Take static scenarios/configurations as baseline, only discuss non-static if static scenarios/configurations are not enough to ensure testing robustness
		- in non-static scenario, it should be ensured that conditions do not very so much that UE would need to change the AI functionality/model
	+ Option 4: Consider non-static scenarios/conditions to ensure minimum performance when AI/ML model change is possible but not signalled by the UE to the NW.
	+ Option 3: Other
* Recommended WF
	+ To be discussed

CMCC: we see the benefit to have non-static condition. It is good to test under non-static condition. Due to time limitation, we can focus on static.

Ericsson: We do not need such discussion. We should identify the scenario and identify whether static or non-static will be used.

Nokia: There are two reasons to consider non-static scenario. 1) To test UE monitoring the change of condition. 2) There are use cases implying using the dynamic conditions.

Vivo: Agree with Option 2. We have discussions per use case. We need to agree with the basic principle and discuss details case by case.

Samsung: Based on our previous agreement both should be use case basis.

Chair: It is not expected that companies continue discussing this topic under the general agenda for AI/ML.

### Sub-topic 1-3

*Static and non-static conditions*

In the previous meeting it was agreed that static and non-static scenarios/configuration could be needed for AI testing, however, no further details were agreed.

Agreement:

* Both static and non-static scenarios/configurations could be needed for AI testing
	+ RAN4 will further discuss how to use them case by case
		- FFS whether to use static scenarios/configurations as baseline.
	+ Refine the definitions of static and non-static scenarios/configurations based on two bullets below
		- Static: channel model and SNR settings are fixed and do not change over the test, specific channel realizations may be dynamic
		- Non-static: Non-static scenarios/configuration can be further considered in application to use cases. The details of models are FFS and may include non-stationary SNR and other conditions.

Some companies brought further inputs on what non-static would be

**Issue 1-3: Non-static conditions definitions**

* Proposals
	+ Option 1: longer term SNR average is changed
	+ Option 2: channel model is changed (e.g. LOS/NLOS, different Doppler shift, etc)
	+ Option 3: Others
* Recommended WF
	+ Further discuss all options

NTT DOCOMO: we think channel model is not changed. We need change the channel parameters.

Qualcomm: The previous conclusion should be applied here. We can continue discussions for this per use case.

### Sub-topic 1-4

*Inference Latency*

Some companies have raised the issue of how to handle inference latency and whether any specific requirements are needed.

**Issue 1-4: Inference latency handling**

* Proposals
	+ Option 1: Postpone discussion until more details on AI/ML procedures are defined in RAN1/2
	+ Option 2: No need for special handling, inference latency will be implicitly captured in the delay of corresponding procedures(e.g. reporting delay) and/or performance requirements(e.g. CQI reporting delay/accuracy)
	+ Option 3: Other proposals
* Recommended WF

Option 2

CMCC: understand the motivation. It is similar as measurement delay. Inference latency is used to check how fast the inference performed. More discussion is needed how to define and verify. We are not OK with option 2.

Apple: inference latency may impact the network configuration. The network needs know when AI model become capable. It is pre-mature to conclude that there is no need. Option 1 is more reasonable.

Nokia: How to control latency is the key. Option 2 is reasonable.

Qualcomm: If option 2 is not agreeable, then the discussions should be based on different use cases.

OPPO: we suggest going with option 2. For model inference like measurement, we did not do the requirement before. How to give reference latency for model interference is hard work.

Samsung: It depends on the use case. Option 2 is understandable.

Intel: We tend to agree with option 2. There is no explicit requirement for latency.

Apple: reporting delay cannot reflect the inference latency given that we have clear definition of latency of inference.

Intel: I have different understanding of inference latency. Apple comment sounds the activation latency/reporting delay.

Apple: we do not need inference latency. We just need reporting delay.

Ericsson: We could have it as starting point. We can further check per use case.

Apple: If there is reporting delay, do we need inference latency? Or they are equivalent.

Intel: Inference latency + measurement delay = reporting delay.

Apple: before UE reports, UE measure and then do inference and report.

Nokia: the example in performance requirement can serve the test purpose of latency of inference.

CMCC: if RAN1 defines the inference latency, how can we handle.

Apple: The reporting will be defined.

Agreement:

* Use the following as the starting point
	+ Inference latency can be implicitly captured in the delay of corresponding procedures (e.g. reporting delay) and/or performance requirements.
	+ Inference latency is part of delay of the corresponding procedure and no separate requirement will be specified for inference latency.
	+ Check the above bullets based on per use case, and if other WGs define the inference latency, RAN4 can revisit the above bullets.

### Sub-topic 1-5

*LCM Requirements*

Several companies proposed to start the discussion on potential LCM requirements even though procedure definition is not yet clear in other WGs

**Issue 1-5: LCM Requirements**

* Proposals
	+ Option 1: RAN4 will define requirements for LCM procedures, candidates are:
		- identification
		- selection
		- activation
		- deactivation
		- switching
		- fallback to non-AI operation
		- performance monitoring
	+ Option 2: Requirements for LCM procedures to be handled similarly to existing RRM requirements wherever possible (e.g. activation delay, switching delay, etc)
	+ Option 3: Postpone discussion until more details from other WGs become available
	+ Option 4: Other proposals
* Recommended WF

Different options to be discussed

Moderator: the options are not exclusive. Option 1 + Option 2.

Vivo: delay requirement is not needed.

Moderator: network will trigger the procedure. It is not pure UE autonomous.

Nokia: this requirement should be per use case specific. The procedure should be general but the value should be use case specific.

Agreement:

* RAN4 considers defining RRM requirements for LCM procedures based on the specific use cases (beam management and positioning), and candidates are:
* identification
* selection
* activation
* deactivation
* switching
* fallback to non-AI operation
* performance monitoring
* Others are not precluded

### Sub-topic 1-6

*Handling of legacy requirements features*

Some companies brought up the issue that the UE should still meet all legacy requirements when using AI/ML enabled features/functionalities

**Issue 1-6: Legacy RRM requirements handling**

* Proposals
	+ Option 1: UE shall meet all legacy RRM requirements (non-AI/ML), even during the AI/ML operation mode.
		- if a legacy feature/requirement is replaced by an AI/ML enabled feature, the legacy requirement would not have to be met anymore
	+ Option 2: Other proposal
* Recommended WF

Option 1

Moderator: when the new requirement is defined, the old one can be skipped.

Intel: it is ambiguous.

Nokia: Similar comment. BM, we have reporting of measurement and prediction, which need to meet the legacy requirements.

Ericsson: we have different understanding of the first bullet. AI/ML may lead to worse performance. Remove the bullet about the replacing the legacy requirements. The legacy requirements would be complementary. We can define some applicability rule.

Agreement:

* UE shall meet all legacy RRM requirements (non-AI/ML) even during the AI/ML operation mode.

### Sub-topic 1-7

Testing goals update

The testing goals were already captured in the TR in RAN4#110, some updates and further agreements are proposed as follows. The options are not exclusive:

**Issue 1-7: Testing goals update**

* Proposals
	+ Option 1: deprioritize option 1 and focus on option 2
	+ Option 2: (R4-2409464):
* 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 3: Updated Option 1 to (R4-2408658):
		- **Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible), can fulfil minimum requirements specified in RAN4.**
	+ Option 4: others, please provide some concrete proposals
* Recommended WF

To be discussed

### Sub-topic 1-8

Testing diagram update/clarification

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

* Proposals
	+ Option 1:
		- Proposal 12 in R4-2408658: RAN4 to agree and clarify in the TS that the reference block diagram in Figure 7.3.2.3-1 in TR38.843 [3] and its description requires further clarifications for the definition of the depicted blocks and links/arrows between them as show in the figure below.



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

To be discussed

### Sub-topic 1-9

*Channels used for training/testing*

The issue of whether synthetic channels would be adequate for performance assessment/testing was brought up and it would be useful to discuss. Whether and how RAN4 could use other channels should also be discussed.

Training here refers to possible training of reference models that used by RAN4.

**Issue 1-9: Training/Test data handling**

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

To be discussed

# 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** |
| [**R4-2407168**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407168.zip) | Korea Testing Laboratory | **Observation 1: It costs a lot to consider all possible combinations between Set A and Set B for the conformance testing.** ***Proposal 1: For the conformance testing, study and specify the reduced or a single combination(s) or between Set A and Set B.*****Observation 2: 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 NW side. Other values, e.g., 256 not precluded.****Observation 3: The number of UE beams can vary by 4 or 8 downlink Rx beams (max number of available beams) per UE panel at UE side. Other values, e.g., 16 not precluded.** **Observation 4: It is essential to provide the spatial correlation for the test setup environment for BM-Case1.** **Observation 5: BM-case1 was evaluated under 3 km/h, with no UE rotation and without a trajectory model. It fits well with the static environments/configurations*****Proposal 2: For BM-case1, deploy the static test environment. (channel model and SNR settings are fixed and do not change over the test, specific channel realizations may be dynamic)*****Observation 6: Since Set A = Set B, the reduced number of beams for Set A and Set B can be an appropriate test environment.** **Observation 7: In the case of Performance when Set A is a subset of Set B, the reduced number of beams for Set B can be adopted as the generalization characteristics of AI/ML.*****Proposal 3: For BM-case2, deploy the static test environment. (channel model and SNR settings are fixed and do not change over the test, specific channel realizations may be dynamic)*****Observation 8: 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 4: The following issues should be considered**** ***Study on the testing BM-Case2 whether or not the inference algorithm is independent of the rotational direction of UE.***
	+ ***Study the test setup on the existing FR2 OTA setup as baseline.***
	+ ***Otherwise, study a test environment considering the rotational direction of UE***

**Observation 9: The testability and interoperability for BM-Case1 and 2 shall have sustainability for the various deployment circumstances of the gNB antenna configuration.*****Proposal 5: 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 the temporal property for BM-Case2.***
* ***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.***
 |
| [**R4-2407234**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407234.zip) | Apple | **Observation 1: The testing environment for Beam Management (BM) case 1 and BM case 2 should replicate conditions that sufficiently capture the correlations of the transmit (Tx) and receive (Rx) beam patterns across the entire spectrum of propagation conditions in both spatial (angles of arrival and departure - AoAs and AoDs) and temporal domains.** **Observation 2: The following conditions will introduce randomness and variations in propagation conditions across both time and spatial domains for the computation of L1-RSRP** 1. **Different AoDs with respect to the Tx antenna array**
2. **Different AoAs with respect to the Rx antenna array**
3. **Different superpositions of {AoA,AoD} pairs**
4. **Fading/Variation in time domain ( different {AoA,AoD} pairs per resolvable delay bin path)**
5. **UE movement (including rotation)**

**Observation 3: During real-world deployment, UE will encounter random radio propagations characterized by variations in both spatial and temporal domains (fading). Testing UE under similar conditions is important to reflect the realities of deployment accurately.****Observation 4: The following questions need to be answered to evaluate the feasibility of the FR2 OTA-based test procedure:** 1. **How can we generate multiple beams from the Set-B and Set-A Tx beams given the limitation of two AoAs?**
2. **What assumption is made regarding Rx beam sweeping? Does the UE utilize a fixed Rx beam, or does it sweep to find the optimal Rx beam?**
3. **How can we simultaneously emulate different AoAs (Rx beam) and AoD (Tx beams)?**
4. **How can we achieve dynamic variation in the AoD domain (Tx beam sweeping) for BM case 2 prediction?**

**Observation 5: From the network operator's perspective, a number of optimal beams need to be known in order to optimize load balancing and make trade-offs in performance and complexity****Observation 6: Ensuring that RSRP accuracy implies beam prediction accuracy is challenging and depends on absolute RSRP accuracy and the RSRP difference between the best beams****Observation 7: When using neural networks for regression, we rely on the Normalized Mean Squared Error (NMSE) criterion for training. It's crucial to ensure that the accuracy of weaker beams is not compromised by the accuracy of the strongest beams****Observation 8: The source of training data for beam management will play a crucial role in AI/ML BM performance and in the generalization performance in real deployment****Observation 9:** **For training data based on real measurements, the** **quality of training data depends on RF impairments, and other noise sources. There is tradeoff between training data quality and generalization performance. With training data collection from the field (across varying SNR conditions), both set-B and set-A beams will be affected ny impairments****Observation 10:** **To guarantee that the UE operates within acceptable margins, it's essential to subject it to various radio conditions and additional conditions for testing and generalization validation in RAN4** **Observation 11:** **For BM use case the identified scenarios and configurations can be initially understood as those reported by UE through capability signaling as part of functionality identification.****Observation 12: The additional conditions for the AI/ML model training (which do not constitute part of UE capability) for the AI/ML-enabled feature/FG can serve as the different scenarios/configuration for defining generalization****Observation 13: Achieving consistency between training and inference by model monitoring could result in delays and increased complexity in model management for BM use case****Observation 14: If multiple models with varying generalization capabilities and requirements for network-side additional conditions are trained by different UE vendors, it would necessitate substantial standardization efforts for BM use case****Observation 15: Current proposals on assistance information for additional conditions and Model Identification only serve the purpose of selecting the appropriate AI/ML model. However, this approach may not be scalable due to considerations of UE implementation complexities and granualtity of conditions/additional conditions. Complexity can increase substantially, especially if condition granularity is fine.*****2.1 Framework for Defining Requirements in Tests for AI/ML*** **Proposal 1**: **RAN4 should study the specification of reference AI/ML models, training procedure, and training data for defining performance requirements for BM use case while considering limitations on model complexity. We provide a reference table for initiating the discussion.*****2.2 Beam Prediction Testability Discussion*****Proposal 2:** **RAN4 should investigate the feasibility of the current FR2 OTA-based test procedure to capture random fading in both spatial and temporal domains, similar to CDL, Uma, etc. Additionally, considering the incorporation of UE rotation can help model randomness from the perspective of the Angle of Arrival (AoA)****Proposal 3: If the current FR2 OTA-based test procedure proves to be not feasible, consider using CDL and Uma channels for testing in the BM AI/ML use case upon TE approval for feasibility** **Proposal 4: For testing the KPIs for BM-Case 1 spatial prediction we propose the test setup and framework described below upon TE vendor approval for the feasibility of this approach.****A diagram of a test  Description automatically generated with medium confidence****A diagram of a diagram of a diagram  Description automatically generated with medium confidence****Proposal 5: For testing the KPI for BM-Case 2 temporal prediction we propose the test set up and framework described below upon TE vendor approval for the feasibility of this approach.** A diagram of a graph  Description automatically generated***2.3 KPIs/Test Metrics for BM Use case*****Proposal 6: We propose RSRP accuracy to be defined as the difference between the predicted RSRP and the genie one (measured) associated with the same Tx beam. Accuracy requirements should apply to all predicted beams which satisfy the pre-defined side conditions, including SNR.**.**Proposal 7: When considering the necessity of additionally testing for beam prediction accuracy, we propose to study the additional information and significance that this test will provide, especially in light of our definition of RSRP accuracy.*****2.4 Measurement accuracy requirements/Training Data Quality for BM*****Proposal 8: When specifying the parameters to generate training data for the UE to train its model, consider the table below as the baseline****Proposal 9:**  **RAN4 to jointly consider the tradeoffs for selecting the source of training and testing data for BM use case.*****2.5 Measurement Acuracy for BM*** **Proposal 10: To investigate performance degradation resulting from measurement accuracy, we propose specifying an impairment model with its associated parameters. This ensures that companies provide comparable simulation results** **Proposal 11: If the outcome of simulating the performance from the impairment model results in degraded performance, then we can consider the following options:*** **Make the AI/ML adaptable to changing SNR conditions, for example to maintain performance the number of set B beams can increase, also the number of K in top-K could increase to compensate for degraded performance.**
* **Change the site conditions, (like SNR, etc) under which the AI/ML operation can be supported. Below a threshold SNR point, legacy procedures should be employed.**
* **If the site conditions significantly change to support AI/ML functionality, then we could consider investigating tighter accuracy requirements**

***2.6 Generalization issues for BM*** **Proposal 12: In the RAN4 core requirement, it is mandated that the consistency or association between Set B beams and Set A beams during both training and inference must be guaranteed. This serves as additional assistance information from the network side for testing UE-sided models during inference and monitoring****Proposal 13: RAN4 should define identified scenarios/configurations associated with the UE capability report of an AI/ML-enabled Feature FG. For defining generalization tests, the additional conditions can serve as the other identified scenarios/configurations for the BM use case****Proposal 14: RAN4 should investigate the feasibility of providing assistance information for the additional conditions to aid generalization and consistency across training and testing when defining requirements. Other additional conditions that are not part of UE capability can be used to define generalization tests****Proposal 15:** **For additional conditions that cannot be shared due to proprietary concerns, RAN4 can explore the feasibility of using a virtual ID to indicate the specific conditions under which a model was trained. This approach would assist in the proper selection of UE models to support generalization. Additionally, RAN4 should identify which additional conditions should be exclusively reserved for generalization tests.*****2.7 Consistency between Training and Inference*** **Proposal 16:  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 17: 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 18: 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 19: 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 20: 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 21: 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 22: 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 23: Investigate the feasibility of training the models with a mixed dataset associated with both network (NW) and UE auxiliary information signals to further enhance the generalizability of the AI/ML model for the beam management case and reduce the number of generalization tests.** |
| [**R4-2407319**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407319.zip) | InterDigital, Inc. | ***Observation 1:*** *The RSRP accuracy requirements are dependent on a minimum Ȇ/Iot level as a side condition.****Proposal 1:*** *The RSRP requirements for baseline test cases to be maintained for beam prediction and data collection.****Observation 2:*** *Defining AIML specific RSRP accuracy requirements for different Ȇ/Iot levels, side conditions may be a good option while maintaining the baseline requirements as fall back.****Proposal 2:*** *Define AIML related extra RSRP accuracy requirements for different side conditions (Ȇ/Iot levels).****Observation 3:*** *If more than one measurement accuracy class is defined, currently there is no measurement quality quantification.****Proposal 3:*** *If specific accuracy measurements are defined in the context of Beam Management AIML, the UE capability should describe it/signal it.****Proposal 4:*** *If specific RSRP accuracy measurements are defined in the context of AIML Beam Management, define a specific new class of side conditions requirements.****Proposal 5:*** *If specific RSRP accuracy measurements are defined in the context of Beam Management AIML, both absolute and relative requirements shall be addressed.* |
| [**R4-2407333**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407333.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:** |
| [**R4-2407367**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407367.zip) | CAICT | **Proposal 1: Suggest to consider the difference between the L1-RSRP of Top-1 predicted beam and the L1-RSRP of the Top-1 actual beam as the definition of RSRP accuracy and further discuss potential test burden with this definition.****Proposal 2: Hold on the discussion on setA and setB consistency until RAN1 has sufficient progress on this.****Proposal 3: RAN4 to study the possibility of reflecting the effect of diverse Tx beams on receiver side through identifying a set of propagation conditions in channel model.****Proposal 4: RAN4 to discuss whether to introduce the testing feasibility and possible methodology for AI/ML beam management into Rel-19 study item on NR FR2 OTA testing enhancement.** |
| [**R4-2407369**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407369.zip) | LG Electronics Inc. | ***Proposal 1***: The predicted RSRP accuracy definition is necessary. we suggest that the definition of predicted RSRP accuracy is difference between predicted RSRP and ideal RSRP in the same beam for the test environment.***Proposal 2***: We suggest that the AWGN channel should be the baseline channel model for test. |
| [**R4-2407497**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407497.zip) | CATT | **Observation 1: The RSRP of predicted beam(s) in the report of inference results can be predicted RSRP or measured RSRP, which depends on whether the beam is configured for corresponding measurement.** **Observation 2: In Option 3, the RSRPs used for comparison are not from the same beam, which may not ensure the RSRP prediction accuracy.** **Observation 3: The difference between Option 2-1 and 2-2/2-3 is that whether a margin shall be considered for beam ID prediction.****Proposal 1: Option 1 and Option 3 can be merged and redefined as follows:** **- Option 1a: The absolute difference between the predicted RSRPs of Top-K predicted beams and the measured RSRPs of the corresponding beams shall be less than X dB,**  **- K should be aligned with RAN1 agreements.**  **- Related measurement accuracy can be considered to determine X.** **Proposal 2: Both KPIs/metrics based on RSRP accuracy and beam ID accuracy should be defined, considering the capability of different types of AI/ML models.** **Proposal 3: Down-selection between option 2-2 and 2-3 which have similar effect as KPIs/metrics can be made after RAN1 have progresses.****Proposal 4: At least Option 2-2/2-3 should be supported if the performance for Option 2-1 and Option 2-2/2-3 are comparable.** |
| [**R4-2407846**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407846.zip) | Xiaomi | **Proposal 1: For test metric option 1, refine the definition by re-using the wording in 38.843:*** **Option 1: Absolute RSRP prediction accuracy is the RSRP difference between the predicted L1-RSRP of Top-1 predicted beam and ideal L1-RSRP of the same beam.**
	+ **FFS definition for ideal L1-RSRP, e.g. real L1-RSRP or measured L1-RSRP.**
	+ **FFS to define relative accuracy requirement.**

**Proposal 2: For test metric option 2, add “FFS definition for Top-1/K genie-aided beam, e.g. real best beam or measured best beam”.****Proposal 3:** **For test metric option 3, refine the definition by re-using the wording in 38.843:*** **Option 3: successful rate for the correct beam prediction which is considered as ideal L1-RSRP among top-1 predicted beams is larger than the ideal L1-RSRP of the Top-1 genie-aided beam – x dB.**
	+ **FFS definition for ideal L1-RSRP, e.g. real L1-RSRP or measured L1-RSRP.**
	+ **FFS definition for Top-1 genie-aided beam, e.g. real best beam or measured best beam**

**Observation 1: From RAN1 agreement, both predicted beam index and predicted RSRP can be included in L1 report at least for BM case-1, similar as legacy.****Observation 2: For option 1, RSRP prediction accuracy focus about the RSRP difference for the same beam.****Observation 3: Both option 2 and 3 can be classified into beam prediction accuracy. Option 2 refers to beam index difference while option 3 refers to RSRP difference between predicted best beam and ideal best beam.** **Observation 4: RSRP prediction accuracy requirement can’t replace beam prediction accuracy.****Proposal 4: RAN4 needs to at least define beam prediction accuracy requirement and discuss which test metric to be chosen, e.g. based on beam index difference or RSRP difference.****Observation 5: Test Data set will be based on synthetic channel model. Therefore, channel model assumption for test will have impact on test data set generation.****Observation 6: The similarity of training data set and inference data set should be guaranteed. Channel model assumed for training and inference test needs to be aligned.** **Proposal 5: RAN4 to apply CDL model as baseline for test data set generation.****Observation 7: For Option 1, for LOS scenario, legacy methodology to decide Gmin and Gmax can be used. However, the beamforming gain range is large.** **Observation 8: For Option 1, for NLOS scenario, it’s more difficult to decide Gmin and Gmax since there are multiple rays. It’s FFS how to decide the combined beamforming gain.****Observation 9: For Option 1, at least for BM case-1, slow time-varying channel can be assumed when deriving ideal L1-RSRP.****Observation 10: Multiple clusters with similar power may have different AOA, AOD.** **Observation 11: For Option 2, for NLOS case, there may be several best TX beam indexes with different AOA and AOD.****Observation 12: For Option 2, for LOS case, the best beam index is unique and TE can know it.****Observation 13: For Option 3, it’s more challenging for TE to know the ideal L1-RSRP for the strongest TX beam. It will include the uncertainty of option 1 and option 2 together.****Proposal 6: RAN4 to discuss whether real best L1-RSRP value or beam index can be known or not:*** **discuss for each test metric respectively**
* **discuss for BM case-1 and BM case-2 respectively**
* **discuss possible condition if real value can be known**

**Proposal 7: RAN4 to discuss whether measured value can be used as reference for ideal value.****Proposal 8: The number of TX beams needs to be defined for both set A and set B. The similar spatial consistency between set A and set B needs to be guaranteed.****Proposal 9: For CDL-D/E channel model, single AOA and single probe can be assumed.****Proposal 10: For CDL-A/B/C channel model, multiple AOA and multi-probe can be assumed. RAN4 further discuss how to simply channel model by reducing cluster number.**  |
| [**R4-2408074**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408074.zip) | MediaTek inc. | **Observation 1**: Only DL Tx beam prediction is in scope and Rx beam prediction is not for both spatial-domain and temporal beam prediction. **Proposal 1**: 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 2**: 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.

**Observation 2**: 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 3**: Either ideal L1-RSPR or measured L1-RSRP could be used for offline training.**Observation 4**: It will be easy for training data collection if we use the ideal L1-RSRP generating from channel emulator.**Observation 5**: 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 6**: Either ideal L1-RSPR or measured L1-RSRP could be used as ground truth for performance evaluation.**Observation 7**: 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 8**: From our simulation results, there is a large inference performance difference between ideal L1-RSPR and measured L1-RSRP.**Proposal 3**: Based on observations 2~8, we suggest RAN4 to investigate the testability of ideal L1-RSRP and measured L1-RSRP in the BM performance test.**Observation 9**: From our simulation results, there is a non-negligible difference between different measurement error modelling when considering measured L1-RSRP.**Proposal 4**: If RAN4 agreed to use measured L1-RSRP in BM test, RAN4 should conclude the error modelling when deriving the requirements.**Observation 10**: Different TE and UE implementation may impact the inference performance of AI/ML BM.**Proposal 5**: 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 11**: The definition of strongest beam in metrics for beam management requirements/tests is not clear. It could be ideal strongest or measured strongest beam.**Proposal 6**: RAN4 should clearly define the ground truth in AI/ML BM testing. To discuss the strongest beam is ideal or measured strongest beam.**Observation 12**: In Option 3, the definition of maximum RSRP among top-K predicted beams is not clear. It could be the maximum ideal L1-RSRP or the maximum predicted L1-RSRP of the top-K predicted beams.**Proposal 7**: RAN4 should clearly define the maximum RSRP among top-K predicted beams in Option 3. To discuss “maximum RSRP” is the maximum ideal L1-RSRP or the maximum predicted L1-RSRP of the top-K predicted beams.**Observation 13**: If the training data samples and the testing data samples are generated under different gNB beam structures, a significant performance loss is observed.**Proposal 8**: RAN4 to carefully discuss on which aspect to consider generalization for AI/ML BM, e.g., DL Tx beam codebook, deployment scenarios, and UE mobility. |
| [**R4-2408179**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408179.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, 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: absolute RSRP accuracy is proposed as the RSRP difference between the predicted RSRP and the genie/measured RSRP with the same ID.******Proposal 4: for Top-K case, it is proposed that RSRP accuracy is applied to all the K beams.*** |
| [**R4-2408292**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408292.zip) | vivo | **Proposal 1: For metrics/KPIs for Beam Management requirements/tests, RSRP accuracy refers to the difference between the predicted L1-RSRP of predicted beam and the ideal L1-RSRP of the same beam.****Proposal 2: The predicted RSRP values corresponding to all of Top-K predicted beams to be reported shall meet the RSRP accuracy requirements (if specified)****Proposal 3: RAN4 could use the framework of FR2 MIMO OTA test as the test framework for AI/ML based BM, by at least considering the following aspects:*** **The number of DL Tx beams that need to be emulated**
* **The number of OTA probes**
* **The range of OTA probes deployment**
* **The channel needs to be reconstructed**
* **The rotation of DUT for emulating movements**
* **The changing of the selected probe/probe power weight/coefficients running in the channel emulator**

Observation 1: Different DL Tx beams are corresponding to different channel environment. For conducted tests, TE needs to emulate different channel coefficients matrix for different DL Tx beams.Observation 2: Consider the UE movements, TE needs to emulate channel coefficients by considering that the AOA/ZOA and doppler of the clusters will be impact for conducted tests.**Proposal 4: For BM case 1, channel emulator to realize the coefficients  for *NB* beams of set B during the total testing time T as****Where K is the number of probes, S is the number of the Tx antenna elements, N is the number of clusters, NB is the number of the DL Tx beams.  is the power weight of k-th probe for the n-th cluster when nB-th DL Tx beam is transmitted in Tx side.  is the polarization matrix.** **Proposal 5: For BM case 1, the probe power weight  can be derived by considering:** * **Same AOA/ZOA of clusters and different number of clusters**
* **Same probe deployment (number and position of probes) for different DL Tx beams**
* **Different probe power weight for different DL Tx beams**

**Proposal 6: For BM case 2, the total testing time T could be divided into NB segments as [T1, T2, …, TNB], where NB is the number of beams of set B. The power weight of probes, channel coefficients and the UE direction need to be considered in every single segment.****Proposal 7: For BM case 2, channel emulator to realize the coefficients  for *NB* beams of set B remain the same as proposal 4 for BM case 1 except velocity becomes vnB. and testing time is TnB** **Where K is the number of probes, S is the number of the Tx antenna elements, N is the number of clusters, NB is the number of the DL Tx beams.  is the power weight of k-th probe for the n-th cluster when nB-th DL Tx beam is transmitted in Tx side.  is the polarization matrix.** **Proposal 8: UE rotation coordinates with the changing of the channel coefficient needs to be considered for BM case 2.****Proposal 9: For BM case 2, the probe power weight  can be derived by considering:** * **Different AOA/ZOA of clusters and different number of clusters**
* **Different or same probe deployment (number and position of probes) for different DL Tx beams**
* **Different probe power weight for different DL Tx beams**
 |
| [**R4-2408604**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408604.zip) | Ericsson | **Observation 1: Training data set herein is the set of information that includes the input (Set B) and output (Set A) of AI/ML model in training phase. Once training data is loaded, a UE is assumed to ‘training ready’.****Proposal 1: RAN4 shall deal with the below two kind of test metrics with separated definitions, and beam information is to be interpreted by RAN1.*** **Option 1: Beam information on predicted Top K beam(s) among a set of beams.**
* **Option 2: 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.**

**Proposal 2: Regarding Top-1 predicted beam, RSRP accuracy could be determined by any of the following:*** **Option1: RSRP accuracy is the difference between the ideal L1-RSRP of the predicted Top-1 beam and the ideal L1-RSRP of the Top-1 genie-aided beam.**
* **Option2: RSRP accuracy is the difference between the predicted Top-1 L1-RSRP and the measured L1-RSRP of the same beam.**
* **Option3: RSRP accuracy is the difference between the predicted L1-RSRP of the predicted Top-1 beam and the ideal L1-RSRP of the same beam.**
* **Option4: RSRP accuracy is the difference between the measured L1-RSRP of the predicted Top-1 beam and the L1-RSRP of the Top-1 genie beam.**

**Among the options, we intend towards Option 3.****Proposal 3: Regarding Top-K predicted beams, RSRP accuracy may be at least one of the following options:*** **Option1: RSRP accuracy is a set of values includes the RSRP accuracy of each beam, and the accuracy of each beam can reuse the definition for RSRP accuracy in K=1 case, then the test is passed if all of the predicted Top-K beams reached the prediction accuracy.**
* **Option2: RSRP accuracy is a RSRP range given the upper bound and lower bound, then the test is passed if all of the predicted Top-K beams are within the range.**

**Between the options, we intend towards Option 1.****Proposal 4: Above RSRP accuracy under discussion is the absolute RSRP accuracy. After the absolute RSRP accuracy is established, relative RSRP accuracy, as described in legacy, shall be determined.****Proposal 5: Measurement error is the difference between the measured RSRP of the Set A/B beam and the ideal RSRP of the Set A/B beam in both training and inference phase. In other words, measurement error occurs at both the input of training and inference phase of AI/ML model.****Proposal 6: Measurement error impact to prediction shall be verified, since the predicted RSRP is always worse than the measured RSRP if measurement error is introduced in training/inference phase of AI/ML model.****Proposal 7: If tightening measurement error under current side condition is difficult from UE implementation perspective, RAN4 can add another side condition for tightened measurement error for enabling AI-ML model.** **Proposal 8: Measurement error impact to prediction shall be verified (e.g. by simulation) to guarantee an acceptable prediction performance. Furthermore, measurement error impact to the predicted beam ID and the predicted RSRP should be studied separately.****Proposal 9: In measurement error, RX beam gain impact may refer to the outcome of the RAN1 study on selecting Rx beam for DL Tx beam prediction evaluations. If a particular beam is chosen when applying AI/ML model, then Rx beam gain error doesn’t contribute error to AI/ML model.****Proposal 10: From RAN4 test perspective,** **consistency between set A and set B is reflected by below aspects:*** **Same NW antenna/beam configurations for set A and set B, and that its configurations don’t change during training and inference.**
* **Same channel model for set A and set B, and that its configurations don’t change during training and inference. (Question: if consistency is valid for a non-static(time-varying) channel for set A and set B?)**

**After agreed, consistency should be taken as a side condition for the requirements.****Proposal 11: RAN4 to use ‘time static’ multipath fading channel model for AI/ML beam prediction tests. RAN4 to further discuss whether ‘time static or time dynamic’ CDL can be used a specific channel model.** **Proposal 12: We can wait for agreements on other issues before agreeing on the options on test environment limitation/requirements which have been listed in the last meeting. Moreover, we may need some simulations to evaluate the requirements.****Proposal 13: No strong view between Option 1 and Option 2, Option 2 at the least is acceptable from training perspective. But regarding Option 2, the question is: how to repeat the test and get same output in case that training data set is updated due to any reason.****Proposal 14: RAN4 to clarify whether training data set covers or doesn’t cover the hardware (analog and digital part) impairment prior to the input of AI/ML model in baseband part at UE.** |
| [**R4-2409001**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409001.zip) | Huawei,HiSilicon | ***Proposal 1:*** For NW-sided beam prediction, the requirement of RSRP accuracy is the measurement accuracy which reuses RAN4 legacy.***Observation 1:*** The test metric in Option 2 is not applicable if the beam measurement accuracy tolerance is not considered.***Proposal 2:*** For UE-sided beam prediction, deprioritize the test metric of beam prediction accuracy in Option 2.***Proposal 3:*** For UE-sided beam prediction, the RSRP accuracy is defined by considering the RSRP difference between the RSRP of the genie-aided Top-1 beam and the measured RSRP of the predicted Top-K.***Proposal 4:*** Before investigating how to set up the test environment, the following questions have to be answered:* How to ensure the consistency between model training and model testing
	+ Whether and how to define a common training/testing dataset.
		- If define, how to capture the different UE implementations and UE behavior when constructing the dataset
	+ Whether UE is expected to collect training dataset and train the model before performing model performance test.
		- If yes, how to resolve the test cost/time issue.
* How to ensure that a UE can pass the test but perform poorly in the field, considering that some parameters used in the test set up which limit the model generalization may totally be different from that in real deployment.

***Proposal 5:*** For test set up in AI-BM, taking the existing FR2 OTA test set up as baseline, any enhancements on top of which should be justified. |
| [**R4-2409470**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409470.zip) | Nokia | **Proposal 1: We support Option1, Option 2 and Option 3 to cover different cases of AI/ML enabled BM use case for both BM-Case1 and BM-Case2. But in Option 3, the value of ‘x’ must be further discussed and must be much lower than the acceptable L1-RSRP measurement accuracy error.****Observation 1:** Different options for performance monitoring under discussion in RAN1 BM might potentially have different core requirements impacts in RAN4.**Proposal 2: RAN4 should consider Core requirements related to performance monitoring for Option 1 (NWside performance monitoring) and Option 2 (UE-assisted performance monitoring), if there is an impact on legacy measurement reporting.****Observation 2:** If legacy measurement accuracy requirements for L1-RSRP are followed during inference phase of AI/ML based BM-Case1, then the performance of L1-RSRP accuracy may be worse than legacy.**Proposal 3: RAN4 should discuss whether it is acceptable to reduce legacy accuracy performance requirement for L1-RSRP in case AI/ML based BM-Case1.****Proposal 4: RAN4 to further consider whether to tighten L1-RSRP measurement accuracy requirements for AI/ML based BM-Case1 to maintain measurement accuracy performance for L1-RSRP.****Proposal 5: RAN4 should consider impacts on latency requirements for known and unknown TCI states for BM-Case1 and BM-Case2.****Observation 3:** If RAN1 design rely only on L1-RSRP measurements for AI/ML based BM use case, then AWGN channel model may be sufficient for testing setup of this use case.**Proposal 6: For AI/ML enabled BM use case, in addition of CDL model, RAN4 should also consider using AWGN channel model for testability aspects.****Proposal 7: For AI/ML enabled BM use case, RAN4 should target to reuse existing test setups/systems and should use number of probes not higher than number of probes used in legacy tests (e.g. multi-Rx).****Proposal 8: UE rotation should be considered in the testing setup for AI/ML enabled BM use case.****Proposal 9: 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 for Generalization Scenarios

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

**Proposal 10: 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 for Scalability Scenarios

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

**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 11: 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.** |
| [**R4-2409684**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409684.zip) | ZTECorporation,Sanechips | **Observation 1: For RSRP accuracy prediction, the RSRP difference is mainly aim to the same beam between the predicted RSRP value and the measured or legacy RSRP values.** **Observation 2: For legacy L1-RSRP accuracy test metric, ideal RSRP is a range not a single value which has the lower and upper bounds.****Observation 3: The relative accuracy is the measured L1-RSRP difference between two RSs.****Proposal 1: The predicted L1-RSRP difference is similar to the legacy, the absolute RSRP accuracy and the relative RSRP accuracy can be:*****Absolute RSRP accuracy= predicted L1-RSRP of beam index i – ideal L1-RSRP of beam index i******Relative RSRP accuracy= predicted L1-RSRP of beam index i – predicted L1-RSRP of beam index n, where the beam index n owns the largest predicted value.*****Observation 4: TE will verify whether the predicted strongest beam ID is the same as strongest measured beam or legacy beam ID. If the strongest beam ID is same as legacy, that means the AI/ML method is better. Otherwise, the test fails.****Proposal 2: Option 2 shall be considered as the performance metrics for beam prediction.****Observation 5: For option 3, there are two different understandings and the different understanding will cause the different outcomes.****Observation 6: Alt.1 does not emphasize whether the strongest RSRP of predicted beam ID is the same with the reference beam ID or not.****Observation 7: Alt.1 may have worse average RSRP prediction accuracy level even if the UE passes the test.****Proposal 3: The possible worse average RSRP prediction accuracy shall be considered if option 3 Alt.1 is chosen to be one of the performance metrics.****Proposal 4: RAN4 shall discuss which alternative shall be considered for option 3 and the understanding for option 3 shall be aligned.****Proposal 5: The description of option 3 shall be refined if Alt.2 is considered and it can be refined as:*****The reference RSRP value of the beam ID corresponding to the maximum predicted RSRP value shall larger than the strongest reference RSRP value minus x dB.*****Observation 8: AWGN model does not have the spatial correlation.****Proposal 6: For AI beam prediction, AWGN model shall not be used and fading channel shall be considered.** |
| [**R4-2409739**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409739.zip) | ROHDE & SCHWARZ | Observation 1: Enhanced IFF has been proven feasible and presents clear advantages with respect to other methodologies for multi-AoA testing.Observation 2: Key capabilities for beam management testing are already supported in current multi-AoA test systems.Observation 3: Enhancing the usage of current testing capabilities (examples given in this paper), ensures more advanced testing environment, which is suitable for AI/ML based beam management.Observation 4: RAN4 should focus on enhancing the usage of current multi-AoA test system, instead of defining a new test system.Observation 5: Defining a new test setup will have a major impact on the requirement definition timeline and system availability.Proposal 1: RAN4 to structure the discussion by defining requirements to the test environment (examples given in this paper).Proposal 2: Use the existing multi-AoA test system as a starting point for AI/ML based beam management test system discussions. |
| [**R4-2409761**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409761.zip) | Samsung | *RAN4 Core Requirements for Supporting NW-sided Model***Observation 1: For NW-sided model for AI-BM, the existing CSI framework for beam configuration (Set A and Set B) will be used.****Observation 2: 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 3: 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 4: From RAN1 perspective, the approaches for overhead reduction could be introduced for Rel-19 AI-BM.****Observation 5: For data collection for NW-side AI/ML model (for BM-Case1 & 2), L1 measurement framework will be used for model inference, and MDT-based measurement/reporting will be used for training.** **Proposal 1: RAN4 shall introduce the necessary core requirement on supporting data collection for NW-side AI/ML model inference/training (for BM-Case1 & 2), by considering:** 1. **Potential enhancement on L1 measurement/reporting for inference: e.g., beam reporting for more than 4 beam in L1 signaling, and overhead reduction;**
2. **Potential enhancement on MDT-based measurement/reporting for training.**

**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: No RAN4 impact is expected for NW-side AI/ML model inference for BM-Case1 & 2, except data collection for model inference.** **Proposal 4: FFS RAN4 impact from NW-side AI/ML model performance monitoring for BM-Case1 & 2 (depending on RAN1/2 input).** **Observation 6: For NW-sided model for AI-BM, UE will neither be informed about any LCM decision, nor be involved in any LCM decision making, except being configured for providing the required measurement/data.****Proposal 5: No RAN4 impact is expected for NW-side AI/ML model LCM for BM-Case1 & 2.** *RAN4 Core Requirements for Supporting UE-sided Model***Observation 7: 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: FFS RAN4 impact from data collection for UE-sided model training (depending on RAN1/2 input).** **Proposal 7: For the testing of UE-sided model inference and monitoring (if any), the consistency of NW-side additional condition across training and inference shall be guaranteed.** **Proposal 8: RAN4 requirement of UE-side AI/ML model inference needs to be specified.** **Proposal 9: RAN4 shall consider the following factors in specifying requirement for UE-side AI/ML model inference: (1) report content of inference results (2) the configuration of inference results reporting.** **Observation 8: 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 UE-sided model LCM, the necessity of RAN4 requirement:** 1. **“Network decision, network-initiated” LCM: RAN4 requirement is needed.**
2. **Other LCM schemes: FFS the necessity of RAN4 requirement (depending on RAN2 input).**

*Summary of Proposals for AI-BM Core Requirement*

|  |  |  |
| --- | --- | --- |
| UE/NW | Operations | Samsung Proposal |
| NW-sided model | Data collection | Proposal 1 (Necessary core requirement for data collection)Proposal 2 (FFS the feasibility part of tightened measurement accuracy in RAN4) |
|  |
| Inference | Proposal 3 (No RAN4 impact expected) |  |
| Perf. monitoring | Proposal 4 (FFS RAN4 impact depending on RAN1/2 input) |  |
| LCM | Proposal 5 (No RAN4 impact expected )  |  |
| UE-sided model | Data collection | Proposal 6 (FFS RAN4 impact for training) |  |
| Additional assistance information | Proposal 7 (Consistency/association shall be guaranteed in RAN4 requirement) |  |
| Inference | Proposal 8/9 (Necessary core requirement is needed) |  |
| Perf. monitoring | Proposal 10 (RAN4 requirement on Type 1 (Option 2) performance monitoring is required.) |  |
| LCM | Proposal 11 (RAN4 requirement for “ Network decision, network-initiated” is required at least) |  |

*Testability Issues for AI-BM***Proposal 12: RAN4 shall study testability study based on existing FR2 OTA chamber systems:** **- Option 1: FR2 SISO OTA chamber, e.g., DFF or IFF (CATR) chamber****- Option 2: FR2 MIMO OTA chamber, i.e., 3D Multi-Probe Anechoic Chamber (MPAC) for FR2 MIMO OTA testing****Proposal 13: The following steps can be followed to emulate the required channel model in OTA chamber:** **Step-1: Determine the required evaluation scenario:** **Step-2: SLS for channel model for all TX beams:** **Step-3: Test signal is mapped over probe(s) in OTA chamber****Proposal 14: FFS the feasibility of DFF chamber and 3D MPAC in terms of:** **- The necessity/feasibility of evaluate RX beam management in OTA chamber;****- The feasibility to generate the required channel model (for certain TX beambook) with limited number of probes.** **Observation 9: By comparing two kinds of existing chambers, the pros, the solution for test signal mapping and expected issues/challenges are provided:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Pros** |  **Cons** | **Test Signal Mapping** | **Issues/Challenges** |
| **Option 1 (FR2 SISO chamber)** | Already used for FR2 RRM testing |  Only 1 AoA emulated, RX beam management impact can’t be tested | Mapping the test signal into tapped delay line signal, which is to be applied on 1 probe | RX beams (at UE) needs to be modeled in the test signal generation, and no possible to test UE RX beam selection/design |
| **Option 2 (FR2 MIMO chamber, i.e., 3D MPAC)** | Possibility for multiple AoAs |  More expensive setup, and exiting 6 AoAs still come from limited spherical area | Mapping requested test signal on 6 probes by TE vendor’s algorithm | Existing 6 AoAs in limited range, which are enough or not for channel model like UMi or CDL needs FFS |

 |
| [**R4-2409770**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409770.zip) | Keysight Technologies UK Ltd | *Observation 1: The AI/ML beam management objective seems primarily targeted for FR2.**Observation 2: The (single probe) IFF test system is the de facto industry standard FR2 OTA test system for UE RF/Demod/RRM test cases requiring just a single AoA.**Observation 3: The Enhanced IFF test system is the de facto industry standard FR2 OTA test system for 2 AoA RRM and Multi-RX test cases while being applicable to all 1 AoA UE RF, demod, and RRM test cases.**Observation 4: The absolute probe locations for the Enhanced IFF test system are not defined and different optimized Enhanced IFF test systems can and have been realized.**Observation 5: The IFF and Enhanced IFF test systems introduce true FF conditions.**Observation 6: The OTA test system for FR2 MIMO OTA conformance test cases utilizes 6 probes with absolute probe locations defined and with the probes implemented with DFF probes.**Observation 7: FR2 MIMO OTA does not introduce true FF conditions as spatial correlation is the predominant requirement.**Observation 8: Only the FR2 MIMO OTA test system can properly and accurately implement CDL channel model.**Observation 9: All FR2 test systems could properly and accurately implement a TDL channel model.**Observation 10: The Enhanced IFF and FR2 MIMO OTA test systems can realize multiple TRPs (with TDL channel models at most) simultaneously.**Observation 11: Only the IFF and Enhanced IFF truly present FF conditions.**Observation 12: The default FR2 MIMO OTA test system cannot present two vastly dislocated TRPs with CDL channel models simultaneously.**Observation 13: Depending on the angular separation of the TRPs, system upgrades of existing FR2 MIMO OTA systems could increase cost&complexity and potentially the chamber size.**Observation 14: Multiple TRPs can be presented to the UE sequentially by rotating the UE but the positioning time in excess of 1s might be too slow.**Observation 15: Multiple TRPs can be presented to the UE sequentially without UE positioner movements with either the Enhanced IFF or FR2 MIMO OTA test system with insignificant delays.**Observation 16: CDL models are realistic channel models for beam management test cases while TDL models which cannot reflect any realism for the beamforming effects at all.**Observation 17: Multiple strong gNB beam candidates for beam selection in multipath propagation environments exist.**Observation 18: Each gNB beam candidates can result in different arrival PAS for the UE reception, potentially impacting also UE beam selection.**Observation 19: Beamforming performed at both link ends (gNB and UE) filters out weak multipath clusters of the channel model, hence applying gNB beamforming simplifies the resulting channel model seen by the UE to a reasonable number of strong clusters.**Observation 20: gNB and UE beam selection impacts resulting cluster-wise fading characteristics.**Observation 21: To properly assess AI/ML algorithms, the complexities of beam dependent fading characteristics and realistic channel models need to be taken into account.**Observation 22: The performance of beam management operations depends strongly on the spatial characteristics of the multipath propagation environment which can be accurately assessed in the FR2 MPAC OTA system.***Proposal 1: It is proposed to limit the testability discussions of AI/ML Beam Management test cases to FR2 only.****Proposal 2: Take the presented FR2 OTA test systems and applicabilities in Table 1 into account for the testability discussions of AI/ML Beam Management.**Table 1: Applicability of existing OTA test systems to testability criteria for AI/ML beam management

|  |  |  |  |
| --- | --- | --- | --- |
| **Applicability Criteria** | **IFF****A diagram of a motor coordinates  Description automatically generated** | **Enhanced IFF****A diagram of a machine  Description automatically generated** | **FR2 MIMO OTA****A grey round object with a green arrow and several blue lights  Description automatically generated with medium confidence** |
| **CDL Channel Models** | No | No | Yes |
| **TDL Channel Models** | Yes | Yes | Yes |
| **Multiple TRPs simultaneously** | No | Yes | Yes |
| **True FF Conditions** | Yes | Yes | Not necessarily |
| **Multiple TRPs with CDL channel models presented simultaneously** | No | No | Requires upgrade whose cost & complexity might be not negligible |
| **Multiple TRPs with TDL channel models presented simultaneously or sequentially at *t*0 and *t*0 + D*T* (without positioner movement)** | No | Yes | Yes |
| **Multiple TRPs with CDL channel models presented sequentially at *t*0 and *t*0 + D*T* (using positioner movement)** | No | No | Yes, but D*T≥*1s might be excessive |
| **Multiple TRPs with at most TDL channel models presented sequentially at *t*0 and *t*0 + D*T* (using positioner movement)** | Yes, but D*T≥*1s might be too slow | Yes, but D*T≥*1s might be too slow | Yes, but D*T≥*1s might be too slow |

**Proposal 3: If more than 1 TRP with spatial separation is required for beam management, limit the maximum number of simultaneous TRPs to 2. The maximum separation of TRPs is FFS.****Proposal 4: For AI/ML beam management requirements, use CDL channel models.****Proposal 5: Infra vendors to clarify the details around gNB beam candidates, including the minimum number of beams, beam steering capabilities/codebook, latency, that should be presented to the UE.****Proposal 6: Consider the FR2 MIMO MPAC OTA test system as the baseline for AI/ML Beam management - DL Tx beam prediction use case.** |

## Open issues summary

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

1. Beam management KPIs

2. RSRP Prediction accuracy

3. Measured/reported value and ground truth - QC

4. Channel model in tests

5. Test setup

6. Datasets

7. Beam consistency

8. Measurement error impact evaluation

9. New UE reporting for network side models

### 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

KPIs and metrics were discussed in the previous meeting without any agreement, RAN4 should continue to discuss what metrics are more appropriate and how they impact requirement definition and testing

**Issue 2-1: Metrics/KPIs for BM**

Proposals

* + Option 1: Use Option 1
	+ Option 2: Use Option 2
	+ Option 3: Use Option 1,2,3 depending on use case
	+ Option 4: Neither Option 1, 2, 3 is appropriate, a new metric is needed
	+ Option 5: Other
* Recommended WF
	+ To be discussed

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

Apple: This discussion is subject to RAN1 discussion. RAN1 does not conclude whether RSRP reporting is included. One possibility is to hold on it until RAN1 finish.

Samsung: discussion is still on-going in RAN1. We should not discuss this issue. Clarification for each option is useful. We do not need discuss it right now.

Nokia: Similar view. It depends on RAN1 conclusion. We may have different options. We support option 3.

OPPO: Share the similar view as Apple. Option 1 need clarifications.

### Sub-topic 2-2

*RSRP Prediction accuracy*

The definition of the prediction accuracy was discussed in the previous meeting, companies were invited to bring further inputs on how this could be defined

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

* Proposals
	+ Option 1: Difference between the predicted RSRP and the measured RSRP of the same Tx beam
	+ Option 2: Difference between the predicted RSRP and the ideal RSRP of the same Tx beam
	+ Option 3: Difference between the predicted RSRP(of the predicted beam) and the ideal RSRP of the beam that should be predicted (genie-aided predicted beam)
	+ Option 4: Difference between the predicted RSRP(of the predicted beam) and the measured RSRP of the beam that should be predicted (genie-aided predicted beam)
	+ Option 5: RSRP difference between the RSRP of the genie-aided Top-1 beam and the measured RSRP of the predicted Top-K
	+ Option 6: others
* Recommended WF
	+ To be discussed

OPPO: besides option 1~5, we have different options. We can copy option 3 and option 4 replacing the predicted RSRP by measured RSRP or ideal RSRP for the predicted beam. How to get the ideal RSRP in RAN4 test? If we could not have ideal RSRP, we need drop options with ideal RSRP.

Mediatek: It is related to ground truth. The question is how to compare RSRP.

Nokia: in legacy requirement, we rely on ideal RSRP. We do not want to drop ideal RSRP.

Apple: we may or may not have predicted RSRP. Shall we wait.

Xiaomi: UE may support both option 1 and 2.

Ericsson: The assumption of question is RSRP should be reported. We do not know why to preclude ideal RSRP.

CATT: to get difference between predicted beam and measured beam, we could not get RSRP from two different beams. Whether to use measured results or ideal results, we can further discuss.

Vivo: Option 1.

ZTE: Option 1 and Option 2 are belong to RSRP prediction accuracy. Option 3 ~5 belong to beam prediction accuracy.

OPPO: for RSRP accuracy, RAN1 already has their definition. Ideal RSRP of predicted beam vs ideal RSRP of the best beam is used in RAN1.

Ericsson: We understand OPPO concern. This is only for beam prediction rather for RSRP accuracy. How can we consider measurement error if we use prediction vs measured ones?

Moderator: how can we know the ideal RSRP? The range of accuracy tolerance compared to ideal RSRP is very large.

Apple: Is the ideal the same as genie?

Moderator: ideal is what UE reports. Genie is for the best beam.

Nokia: Either align with the definition of legacy or enhance the accuracy requirements.

Apple: We have reported RSRP of UE which is based on prediction. Once UE predicts the beam, the UE can do measurement afterwards. We can assume very high SNR. We can get the genie RSRP.

### Sub-topic 2-3

*Measured/reported value and ground truth*

The ground truth is the ideal value that the UE should report

**Issue 2-3: Reported measurements and ground truth**

* Issues raised:
	+ Option 1: UE reported measurement of the same Tx beam taken at a different time(e.g. before UE doing prediction) should be taken as the ground truth
	+ Option 2: Ground truth cannot be determined
	+ Option 3: consider a range for the ground truth based on UE antenna gain(similar to the reporting range for current FR2 measurement accuracy tests)
	+ Option 4: others
* Recommended WF
	+ To be discussed

Mediatek: for measurement, it could be measurement with noise on the condition with high SNR. Measurement under high SNR could be ground truth.

Nokia: Option 2 does not make sense. How to define the ground truth depends on the test design.

Qualcomm: before coffee break, we have already had scheme to show what the ground truth is. UE just reports the measured one if option 1 is used. Option 1 needs add the additional condition that the channel condition should be kept the same.

Qualcomm: for option 1, opposite is also true. You can just report the predicted rather than measurement.

Intel: Ground truth is more useful for training. What is the ground truth here?

Moderator: ground truth is what UE should ideally predict. Ideal has 13dB range.

Apple: We can take Mediatek comment. We can define the ground truth as the measurement in the high SNR.

Vivo: this approach does not work. RF impairment and other aspects will be included. What accuracy do you expect?

Nokia: Since we are now deciding to use measurement as basis, it may be one of candidate solutions.

Samsung: For measurement, you have measurement accuracy. Apple proposal is the reported value under high SNR.

NTT DOCOMO: if the ground truth is defined by measurement under high SNR, how to consider the ground truth for low SNR case.

Moderator: the RSRP should be the same in high and low SNR.

Samsung: even if UE is in the same location, the predicted beam comes from the other direction.

Agreement:

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

### Sub-topic 2-4

*Channel models in tests*

**Issue 2-4: Channel models**

* Proposals
	+ Option 1: CDL
	+ Option 2: TDL
	+ Option 3: AWGN
	+ Option 4: others
* Recommended WF
	+ Option 1

R&S: need understand the value of using CDL. There is complexity.

Qualcomm: we provide the reasons in our paper. Compared to TDL, CDL is with multiple AoAs. You can have separate Tx beam gains and Rx beam gain. One AoD + one AoA, only the channel gain can be changed. The channel during SSB burst will not be changed with TDL which make test be too simple.

R&S: as we mentioned, there is still potential. There are many dimensions in the current system. How can we play CDL in the current test scenario? We need think about the feasibility.

Nokia: the input of model is RSRP. Why we need CDL model. RSRP is the power of signal. Path fading is not critical. What is the reason to consider CDL modelling?

Apple: To Nokia, tend to agree with Qualcomm. We need set B to predict. With TDL, we only have one AoA. We need the input from TE vendor about the feasibility about the CDL. If not we can use the other solution. Currently we cannot mimic the beam sweeping.

Samsung: What kinds of test setup do you want to use? If using Rel-15, we can only use TDL. If not, MIMO OTA would be one option.

Intel: If going with CDL model, we end up with no requirement in this release. It will take long time. The existing test setup can work.

Qualcomm: to Samsung, we need start with what we want to test. Maybe it takes more time to do CDL. It may be useful. We should avoid meaningless testing.

R&S: At least the existing test setup has capability for such exercise. We can rotate UE and change the direction.

Qualcomm: We need understand what we are trying to test and emulate. We are trying to test with signal from different AoAs. MIMO OTA test setup would be used. How to model Tx beam on the sweeping. From R&S paper, we do not know whether the Tx beam can sweep and whether the received power from different AoAs can be different.

Ericsson: How to use the existing test setup to mimic CDL channel is the next step. We believe test setup can be configured to realize CDL channel. We can first decide the direction and then discuss how to realize.

Apple: The key question is how quickly the beam can be rotated. The beam should be rotate within us.

R&S: For this scenario, we can just switch the probes.

Keysight: we have already known how to implement CDL, which is established system. MIMO OTA system can be extended. Complexity and cost can be done depending on how many AoAs.

R&S: Regarding reusing MIMO OTA, it is very difficult to scale in terms of time, cost and very limited use. We would like to use one AoA or two AoA test setup.

Qualcomm: to R&S, everything you mentioned is to adjust power.

### Sub-topic 2-5

*Test setup*

A list of test setup/needs should be created in order to see what kind of test setup is needed and what is feasible

**Issue 2-5: Test setup needs**

* Proposals
	+ Option 1:
		- number of Tx beams
		- AoA, AoD
		- Propagation conditions (including need for LoS/NLoS)
		- UE rotation yes/no
	+ Option 2:
		- other parameters
* Recommended WF
	+ To be discussed

### Sub-topic 2-6

*Data sets*

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

* 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

### Sub-topic 2-7

*Beamforming consistency*

multiple companies brought up the need to discuss the consistency between set A and set B, otherwise it is not expected that beam prediction would work

**Issue 2-7: Consistency**

* Proposals
	+ Option 1: consistency between set A and set B is reflected by below aspects:
		- Same NW antenna/beam configurations for set A and set B, and that its configurations don’t change during training and inference.
		- Same channel model for set A and set B, and that its configurations don’t change during training and inference. (Question: if consistency is valid for a non-static(time-varying) channel for set A and set B?)
	+ Option 2: Consistency should be defined in a different way
	+ Option 3: no need for any consistency definition
* Recommended WF
	+ To be discussed

### Sub-topic 2-8

*Measurement error impact*

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

**Issue 2-8: Measurement error impact**

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

### Sub-topic 2-9

*UE reporting for network side models*

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

**Issue 2-9: UE reporting for network side models**

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

# 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** |
| [**R4-2407235**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407235.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** |
| [**R4-2407498**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407498.zip) | CATT | **Proposal 1: Postpone the discussions on KPIs/metrics for case 1 until RAN4 decide to define accuracy requirements for case 1.** **Proposal 2: Do not further discuss KPIs/metrics for case 3a/3b since RAN4 will not define positioning accuracy requirements.** **Observation 1: RAN1 have been discussing case 2a/2b, though they are second priority.** **Proposal 2: For case 2a,****- For legacy measurement quantities, the existing accuracy requirements in positioning WI can be the starting point.****- For new measurement quantities introduced by other WGs, RAN4 can further discuss the accuracy requirements.** **Proposal 2: For case 2b,****- For legacy measurement quantities, the existing accuracy requirements in positioning WI can be reused.** **- For new measurement quantities introduced by other WGs, RAN4 can further discuss the accuracy requirements.** **- RAN4 do not define accuracy requirements for the UE position outputted by AI/ML models in LMF side.**  |
| [**R4-2407836**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407836.zip) | Xiaomi | **Observation 1: If no requirements defined for AI Pos cases, it would lead some inconsistent behaviour and requirements on PRS measurements when AI based positioning fallback to non-AI mode.** ***Proposal 1: In order to guarantee the consistent UE behaviour and requirements, RAN4 should define accuracy requirements for AI Pos case 1 and case 3a/3b.*****Observation 2: The legacy PRS measurement requirements up to Rel18[5] are based on the path based measurement.*****Proposal 2: If RAN1 using sample-based measurement as model input for AI Pos, the existing accuracy requirements (e.g. PRS RSRPP) shall be restudied at least.******Proposal 3: AI performance monitoring can be used for the post deployment verification directly.*****Observation 3: For AI model performance monitoring, the proper threshold used to justify AI model validity can be studied by RAN4.****Observation 4: For Case 1 if AI model performance monitoring handled by UE, the requirements on model evaluation period can be defined in RAN4.*****Proposal 4: RAN4 shall study the following necessary requirements for AI pos model monitoring. The specification impacts on TS38.133 or other 3GPP spec can be FFS also.**** ***Monitoring hypothesis and threshold***
* ***Performance evaluation period***
 |
| [**R4-2408176**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408176.zip) | CMCC | ***Proposal 1: for case 3a, it is proposed to define accuracy requirements for the measurement reported by gNB, and legacy gNB based positioning requirements can be used 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: it is proposed to consider CIR/PDP as new measurement to define requirements, considering that CIR/PDP is input of model which has impact on the performance of model.******Proposal 3: 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.*** |
| [**R4-2408293**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408293.zip) | vivo | ***Observation 1: Potential report schemes need to be introduced in performance monitoring procedures according to RAN1 agreements and RAN4 requirements for delay/accuracy may need to be considered in different options.******Observation 2: Delay/accuracy requirement for performance monitoring procedure in case 1 may need to be defined and more progress from RAN1/2 are needed for detailed further discussing.******Observation 3: Accuracy requirement for UE position in case 1 are to be defined for RAN4 if the monitoring procedure is performed at LMF side based on the current RAN1 agreement.*** |
| [**R4-2409002**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409002.zip) | Huawei,HiSilicon | ***Proposal 1:*** For UE-sided direct AI/ML positioning (Case 1), positioning accuracy is not testable.***Observation 1:*** For UE-assisted positioning with LMF-side positioning (Case 2b), the relationship between measurement accuracy and positioning accuracy is unavailable, which has an impact on the test requirement definition.***Proposal 2:*** If RAN4 studies how to test Case 2b in AIML for positioning, the relationship between measurement accuracy provided by UE and the eventual positioning accuracy at LMF needs to be investigated firstly. |
| [**R4-2409579**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409579.zip) | Ericsson | **Proposal 1**: 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. |
| [**R4-2409648**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409648.zip) | Nokia | **Observation 1: If an LCM action is required and it is not taken in a timely manner, the performance degradation for AI/ML enabled Positioning use case may be degraded to undesirable level.****Proposal 1: RAN4 to define the time latency limit on UE’s LCM actions when an LCM procedure is indicated by network.** **Proposal 2: 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).****Proposal 3: 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).****Proposal 4: RAN4 to discuss whether any measurement accuracy requirement is needed for at least the types of time domain channel measurements supported by RAN1 in Case 2b.**  |
| [**R4-2409685**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409685.zip) | ZTECorporation,Sanechips | **Observation 1: There is no need for RAN4 to consider model inference for case 2b since the AI/ML model is deployed at LMF side which the output UE location on the LMF side depends on the network implementation.****Proposal 1: RAN4 shall not define the accuracy requirements for case 2b, it is up to network implementation.****Observation 2: In legacy, there are two kinds of location request: UE-triggered location request and NW-triggered location request. The latest agreement contains the implicit location request which is the NW triggers the location request, otherwise no reporting scheme is needed.****Proposal 2: RAN4 shall wait for RAN1 agreements on defining reporting scheme.****Observation 3: For AI/ML based positioning, the main difference compared to the legacy is that the AI model resides within the LMF. The accuracy of position prediction conducted by the LMF depends on the measurements reported by UE.****Proposal 3: The intermediate features such as RSTD, RSRP etc. or some enhancements shall be considered.****Observation 4: From request to reporting time, there are two periods of time which are PRS measurement and model inference. If there is no limitation on these two periods, LMF would take a long time to receive an inference outcome or the intermediate features. It has the impact on the performance accuracy since the accuracy of model output will also decrease over time.****Proposal 4: RAN4 shall define the delay requirements from LMF requesting to UE reporting.****Proposal 5: ToA shall be the intermediate feature when studying the accuracy requirements for case 2a.****Observation 5: RAN1 just agreed at least LOS/NLOS indicator is supported for reporting.****Proposal 6: LOS/NLOS could be the intermediate feature to be reported and RAN4 shall consider how to define the requirements for LOS/NLOS.**  |

## Open issues summary

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

1. Requirements for case 1
2. Requirements for case 2a
3. Requirements for case 2b
4. Requirements for other reported metrics

### Sub-topic 3-1

*Requirements for case 1*

In the previous meeting it was agreed to postpone the discussion until a reporting scheme, if defined, is clear. Some companies are proposing not to define any requirements for this case

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

* Proposals
	+ Option 1: RAN4 should not define requirements for case 1
	+ Option 2: RAN4 should postpone the discussion until other WG conclude on defining a reporting scheme or not
	+ Option 3: Others
* Recommended WF

To be discussed

Note: currently there are no requirements for UE based positioning

### Sub-topic 3-2

*Requirements for case 2a*

Requirements for case 2a have been deprioritized, however, some companies are proposing to discuss this anyway.

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

* Proposals
	+ Option 1: RAN4 should not define any positioning accuracy requirements because positioning is LMF based
	+ Option 2: RAN4 should continue to discuss how to define requirements for case 2a, companies should bring more concrete proposals in future meetings
	+ Option 3: Postpone discussion until other groups make more progress
	+ Option 4: others
* Recommended WF
	+ To be discussed, options are not exclusive

### Sub-topic 3-3

*Requirements for case 2b*

Requirements for case 2b have been deprioritized, however, some companies are proposing to discuss this anyway.

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

* Proposals
	+ Option 1: RAN4 should not define any positioning accuracy requirements because positioning is LMF based
	+ Option 2: RAN4 should continue to discuss how to define requirements for case 2b, companies should bring more concrete proposals in future meetings
	+ Option 3: Postpone discussion until other groups make more progress
	+ Option 4: others
* Recommended WF
	+ To be discussed, options are not exclusive

### Sub-topic 3-4

*Requirements for other reported metrics*

Some companies are proposing to already start discussing how to define requirements for possible reported metrics such ToA, LoS/NLoS, etc. These might be introduced for cases 2a/2b, 3a/3b

**Issue 3-4: Requirements for reported metrics**

* Proposals
	+ Option 1: RAN4 to postpone discussion until reported metrics become clear in other groups
	+ Option 2: RAN4 to already start the discussion on how to define requirements for LoS/NLoS indicator
	+ Option 3: RAN4 to start discussing how to define requirements for other reported values, e.g. CIR/PDP, ToA,
	+ Option 4: others
* Recommended WF
	+ To be discussed, options are not necessarily exclusive

# 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** |
| [**R4-2407236**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407236.zip) | Apple | **Observation 1: There are numerous critical issues that must be addressed to assess the feasibility of option 3.****Observation 2: In the discussion on RAN4-RAN1 coordination, it has been suggested that RAN-1 option 1, fully standardizing the full decoder, is similar to option 3 of RAN4 for fully specifying the test decoder. However, in our opinion, the two options are not similar. This is because the fully specified test decoder would only be applicable to testing, allowing infrastructure vendors the freedom to implement their own test decoder. In contrast, in RAN 1 option 1, NW vendors would be required to implement the standardized decoder.****Observation 3: A test decoder trained individually with the UE’s encoder has less representation learning and they are less flexible. Moreover, performance results would be better since the UE encoder is jointly trained with a “matched” test decoder. On the contrary a test decoder trained with a mixed dataset has better representation learning and it is more flexible but its performance when paired with a particular Encoder could be worse. Performance results with the mixed dataset trained test decoder would be degraded but the test decoder would be more flexible (better generalization ability)** **Proposal 1: For aligning the results between companies and for assessing the performance differences between different schemes, we propose that companies report along with simulation results, the method of obtaining the encoder input: genie channel estimates versus real channel estimation, and details of the impairment model (SNR, channel estimation error etc).****Proposal 2: To assess the feasibility of option 3, RAN4 should address several key considerations. These include:** * **With training type 1 it would be challenging to ensure the fully specified 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 not be constrained to match test decoder implementation) The testing procedure should provide some confidence that if UE passes the test, it will likely exhibit similar performance in the field across tailored NW decoder implementations.**

**Proposal 3: We propose to train the test decoder with a database collected across multiple UE vendors to enhance the generalizability of the test decoder for testing multiple UE vendors. Each vendor will pre-train the test decoder with the database of labeled data sets. Therefore, we propose to employ Training Collaboration type 3 for developing the test decoder.****Proposal 4: To fully specify the Option 3 test decoder and derive RAN4 performance requirements, follow the steps outlined in the attached flowchart procedure.****A diagram of a process  Description automatically generated****Fig. 3: Proposed Flowchart for Standardizing the test Decoder and deriving performance requirement****Proposal 5: For generating data for training the pair of UE encoder/Test Decoder, use a system-level simulator. The proposed baseline system simulation assumptions are shown in Table 2** **Table 2: Proposed Baseline System Level Simulation assumptions for generating training data for Option 3 Test decoder**

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform | FDD, OFDM |
| Multiple access | OFDMA |
| Scenario | Dense Urban (Macro only) Urban MacroMacro Cell |
| Frequency Range | FR1 only, 2GHz, 4GHz  |
| Inter-BS distance | (1)200m, (2) 750 Urban macro or (3) 1732m Macro cell  |
| Channel model         | According to TR 38.901 |
| Antenna setup and port layouts at gNB | Companies need to report which option(s) are used between- 16 ports: (8,4,2,1,1,2,4), (dH,dV) = (0.5, 0.8)λ- 8 ports: (4,4,2,1,1,2,2), (dH,dV) = (0.5, 0.8)λ |
| Antenna setup and port layouts at UE | 2RX: (1,1,2,1,1,1,1), (dH,dV) = (0.5, 0.5)λ for (rank 1,2) |
| BS Tx power | 41 dBm for 10MHz, 44dBm for 20MHz, 47dBm for 40MHz |
| BS antenna height | 25m |
| UE antenna height & gain | Follow TR36.873 |
| UE receiver noise figure | 9dB |
| Modulation | Up to 256QAM |
| Numerology | Slot/non-slot | 14 OFDM symbol slot |
| SCS | 15kHz for 2GHz, 30kHz for 4GHz |
| Channel bandwidth | 10 MHz for 15kHz40 MHz for 30kHz SCS  |
| Frame structure | Slot Format 0 (all downlink) for all slots |
| Feedback assumption | Realistic |
| Channel estimation          | Realistic. Up to companies to choose the error modeling method for realistic channel estimation  |
| Evaluation Metric | NMSE, SGCS  |

**Proposal 6: For deriving the performance requirements (after the specification of test decoder) for the encoder design we should use a Link Level Simulator that reflect the testing conditions at TE.** **We can discuss LLS assumptions at a later stage based on alignment with SLS** **Proposal 7: 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** |
| [**R4-2407334**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407334.zip) | Qualcomm Incorporated | **Proposal 1: Consider the following evaluation parameters:**

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform  | FDD OFDM  |
| Bandwidth | 20MHz |
| Subcarrier spacing | 15kHz  |
| Nt | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ |
| Nr | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ |
| Channel model | CDL-C or TDL-A |
| Doppler spread | 10Hz doppler |
| Delay spread | 30ns  |
| Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) or ideal DL channel estimation |
| Rank per UE | Rank 1 |
| Latent message size | Use power of 2, choose from 32, 64, 128, 256 and 512 bits. |

**Channel model requires more discussion in RAN4, we analyze both options in the following:*** **CDL-C: RAN1 already use it for evaluation and it is captured in TR. RAN1 have done much system level evaluation and CDL model aligns better with RAN1 system level evaluation. However, it is different from RAN4 CSI test setup, and based on R19 channel model SID discussion, emulation for such channel in RAN4 environment could be infeasible and simplification is needed.**
* **TDL-A: It’s a common channel condition used in RAN4 test setup and RAN4 evaluations. No issue for emulation feasibility. However, RAN1 never evaluated channel conditions similar to TDL channels before, and more efforts are needed for RAN4 evaluation if we choose TDL channels.**

**Proposal 2: We propose a set of test decoder model parameters in Table 2 and depict it in Figure 1. Although the proposal is for test decoder, we include the paired encoder (Figure 1) to better represent the overall structure proposal.**Input(nSBxnTx)MLPFlattenOutput layerLatent messageLatent messageInput layerReshapeMLPOutput(nSBxnTx)**Table 2: Test decoder parameter proposal**

|  |  |  |
| --- | --- | --- |
| Category | Parameter | Proposal |
| Model architecture parameters (orange) Assumption: encoder model is the mirror of decoder model | Model type | MLP  |
| Model depth | Three linear layers (with one activation function) |
| Layer type/size | 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.

Note: 2nd linear layer, Activation function, 3rd linear layer are shown as “MLP” is Figure 1. |
| Fixed point representation | TBD |
| Format of input to encoder/output of decoder | TBD |
| Quantization method for the encoder output (encoder) | Scalar quantizer, with 2 bits per dimension |
| Training related parameters | Loss function | SGCS |
| Training dataset | Encoder input dataset should cover all the contributing companies’ encoder input data |

**Proposal 3: In the feasibility study of demonstrating that deriving a fully specified test decoder is feasible, RAN4 first discuss to decide model type for the test decoder.****Proposal 4: 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/decoder pair**
	+ **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/decoder as a reference encoder/decoder pair 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 5: 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 4) 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/decoder pair): 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.**

**Observation 1: RAN4 can consider Table 2 and Figure 1 for derivation of the highlighted entries in the below table (copy of Table 3) 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
 | 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/pair 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 6: RAN4 continues to study a feasible (from test repeatability perspective) and implementation/specification friendly solution to option 4.** |
| [**R4-2407368**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407368.zip) | CAICT | **Proposal 1: For option 3, a unified design of test decoder and reference decoder is preferred.****Proposal 2: For option 3, suggest to consider the possibility of introducing reference encoder for defining requirement requires.****Proposal 3: For option 4, whether to consider test decoder verification in RAN4 requires further discussion.****Proposal 4: If test decoder verification is considered in RAN4, a pair or pairs of reference encoder and reference decoder is suggested to be specified with model structure at least.****Proposal 5: Suggest to consider following aspects on the flow chart provided in R4-2405653.*** **Add a step, “identify target cases with specific test conditions” before step 1.**
* **For step 1, suggest to limit the scope of potential model types. Transformer-based could be prioritized.**
* **Add a step, “Define evaluation methodology” before step 4.**
* **Suggest to align the understanding of step 5 on, whether only a single model architecture would be determined for 2-sided case at least for this release.**
* **Discuss the necessity of step 8 and 9.**
 |
| [**R4-2407499**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407499.zip) | CATT | **Proposal 1: RAN4 discuss the following aspects in high priority for standardization/simulation for Option 3:** **- 1st priority: Model type (transformer / CNN / MLP).** **- 2nd priority: Constraint conditions (Maximum memory [10 M] / complexity [15 M], etc.).** **- 3rd priority: Evaluation metrics (SGCS, etc.).** **The impacts of other aspects can be further studied in iterations.** **Proposal 2: The following simulation assumptions can be the starting point:** Table 1: Baseline Link Level Simulation assumptions for AI/ML based CSI feedback enhancement evaluations

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform  | FDD, OFDM  |
| Carrier frequency | 2GHz |
| Bandwidth | 10MHz |
| Subcarrier spacing | 15kHz for 2GHz |
| Nt | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ |
| Nr | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ |
| Channel model (Delay spread) | CDL-C (300 ns) as baseline, CDL-A (30 ns) as optional. |
| UE speed | 3kmhr |
| Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline.Ideal DL channel estimation is optionally taken into the baseline of evaluation methodology for the purpose of calibration and/or comparing intermediate results (e.g., accuracy of AI/ML output CSI, etc.). Up to companies to report whether/how ideal channel is used in the dataset construction and performance evaluation/inference. |
| Rank per UE | Rank 1.  |
| Evaluation metric(s) | SGCS |
| Note: the baseline EVM is used to compare the performance with the benchmark release, while the AI/ML related parameters (e.g., dataset construction, generalization verification, and AI/ML related metrics) can be of additional/different assumptions. The conclusions for the use cases in the SI should be drawn based on generalization verification over potentially multiple scenarios/configurations. |

**Other factors are open to discuss.** **Observation 1: Some unimportant conditions identified in Option 3 discussions can be relaxed or excluded from RAN spec, which, in our understanding, is Option 4.** **Proposal 3: RAN4 focus on feasibility of Option 3 due to limited time before checkpoint in Sept, and come back to study Option 4 after Option 3 is confirmed feasible, if possible, to leave some implementation flexibility to TE vendors.**  |
| [**R4-2407847**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2407847.zip) | Xiaomi | **Observation 1: For option 4, RAN4 agrees that reference model will be partially specified. However, there is no clear definition what is partially specified for a model.****Observation 2: Input/output of model, i.e. dataset or dataset format, can also be part of reference model.****Proposal 1: Before discussing whether the reference model definition is the same for both option 3 and option 4, RAN4 needs to discuss what is partially specified for a reference model for option 4 first.****Proposal 2: RAN1 is also discussing interoperation issue, options in RAN1 can be classified into options in RAN4:*** **Option 1 in RAN1 is quite similar as Option 3 in RAN4**
* **Option 5 in RAN1 is some kind of option 1/2 in RAN4**
* **Option 2~4 in RAN1 can be classified into option 4 in RAN4**

**Observation 3: For option 3, it’s challenge to align all the parameters. Besides, if several reference models need to be aligned for different scenarios, the specification effort is quite high.****Observation 4: From RAN1 conclusion, fully specified model has limited performance in the field compared to other options. It also requires high specification effort from RAN1 perspective.****Proposal 3: For option 3, RAN4 to take RAN1’s observation into consideration.****Proposal 4: For option 4, RAN4 to discuss test feasibility issue for:*** ***Option 4: Standardized data / dataset format + Dataset exchange between NW-side and UE-side***
 |
| [**R4-2408178**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408178.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.*** |
| [**R4-2408294**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408294.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.** **Proposal 1: RAN4 to define reference model for defining performance requirements for both one-sided model and two-sided model.****Proposal 2: Reference encoder and decoder are needed for requirement derivation, and could largely be reused for test decoder derivation for both Option 3 and Option 4. RAN4 to work on reference encoder and decoder firstly. Later to discuss what will be put in the spec.****Proposal 3: The definition of reference encoder/decoder is: the encoder/decoder used in RAN4 discussions at least for simulation alignment/requirement derivation, test decoder derivation and/or test decoder verification. It could be documented (in TR, WF, etc) or captured in the specifications as necessary.****Proposal 4: After the alignment of reference encoder and decoder, what will be put in the spec for Option 3 and Option 4 is listed in the following.*** **Option 3: Test decoder (+ reference encoder model structure), and channel generation method are in the spec;**
* **Option 4a: Reference encoder + test decoder model structure, and channel generation method are in the spec;**
* **Option 4b: Test decoder model structure + reference encoder model structure, and dataset (PMI and corresponding channel) are in the spec;**

**Proposal 5: One test decoder could be used for one test case or multiple test case.****Proposal 6: Different reference encoder may be defined for different requirement or test case.****Proposal 7: As seen in Figure 2.3-1, the reference/test encoder/decoder may be aligned through the following procedures*** **Step 1: Align the dataset containing only channel information.**
* **Step 2: Determine the model hyperparameters that need to be aligned.**
* **Step 3: Define the evaluation method for model complexity and performance.**
* **Step 4: The best model structure(s) may be selected based on the aligned evaluation method, through the simulations using the aligned dataset.**
* **Step 5: Based on the aligned model structure, the specific parameters of the reference model would be merged from companies**

**Proposal 8: Take into consideration the parameters to be aligned for Option 3 and Option 4 in Table 2.3-1.****Proposal 9: The suggested model structures in Figure 2.3-2 to 2.3-5 for test decoder/encoder could be used as a starting point.** **Proposal 10: Using the mixed dataset for model training, including the mixing of TDL, CDL and UMa, while using the TDL dataset for RAN4 tests. Other mixing rules are not precluded.****Observation 2: From initial results for field test, the generalization performance of AI/ML model trained by UMa simulation data on field data seems acceptable, which has similar performance as eType II codebook. The generalization performance of AI/ML model trained by CDL simulation data on field data is worse than AI/ML model trained by UMa simulation data.****Proposal 11: “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 12: Take into consideration the summary of 4 options for testing of 2-sided model in Table 2.3-6.****Proposal 13:** **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 14: Since Monitoring is still under discussion in RAN1, RAN4 should wait for further progress of RAN1.** |
| [**R4-2408492**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408492.zip) | Ericsson | [Proposal 1 A reference model is a model agreed in RAN4, but not necessarily used for testing and not mandatory in any implementation.](#_Toc166491822)[ For option 3, a reference model (e.g., a reference encoder) could be standardized and used for verification of implementation of a standardized test model (e.g. test decoder) or an actual proprietary model (e.g. decoder).](#_Toc166491823)[ For option 4, a reference model (e.g. a reference encoder) could be standardized as a means of capturing the latent space. A test decoder can be developed based on the reference encoder and a suitable training dataset.](#_Toc166491824)[ For both 2-sided and 1-sided AI functionality, it may be necessary to agree on a reference model in order that the specified performance requirement at an agreed complexity and implementation benchmark can be derived. In this case, the reference model may not need to be standardized it could be documented in a TR (or even no documented at all).](#_Toc166491825)[Proposal 2 The test decoder needs to have a latent space and be trained with a dataset wide enough to cover real expected scenarios. Otherwise, it may not be possible to create real encoder/decoders that both pass the test and provide useful performance in real operation.](#_Toc166491826)[Proposal 3 The table of model architecture is provided in section 2.2 as input to the discussion.](#_Toc166491827)[Proposal 4 It is useful but not essential to report or agree training parameters to enable alignment. Section 2.2 provides a table of training parameters for discussion.](#_Toc166491828)[Proposal 5 Use a dataset obtained from RAN1 system simulations as a starting point for comparing proposed test decoders.](#_Toc166491829)[Proposal 6 For option 4, the latent space needs to be standardized](#_Toc166491830)[Proposal 7 For option 4, standardize a reference encoder in order to capture the latent space](#_Toc166491831) |
| [**R4-2408616**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408616.zip) | Intel Corporation | **Proposal #1: Further discuss and agree on the procedure to conduct analysis of Option 3/4 and on the principles to select the decoder for specification.****Proposal #2: Further align the AI/ML test methodology with RAN1's conclusions on inter-vendor training collaboration for AI/ML-based CSI compression.****Proposal #3: Adopt the following definition of Reference decoder/encoder*****Reference decoder/encoder: The decoder/encoder model used to define the minimum performance requirements. The Reference decoder is identical to the Test decoder.*****Proposal #4 Further discuss and define the upper bound complexity including the number of computations and number of parameters for test/reference encoders/decoders.****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.** |
| [**R4-2409003**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409003.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. ***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. |
| [**R4-2409087**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409087.zip) | ZTE Corporation, Sanechips | ***Observation 1. CDL channel model is still discussed for Rel-19 work item in the RAN plenary.******Observation 2. Further clarification is needed on whether the latent message is before or after quantization if RAN4 is consider standardizing the latent message.******Proposal 1. The encoder/decoder used in RAN4 discussions for simulation alignment/requirement derivation. It could be documented or captured in the specifications as necessary.******Proposal 2. Table 6.2.1-2 in TS 38.843 can be used as a reference, more detail parameters need to align with RAN4 existing configuration.******Proposal 3. Considering the following table parameters for model architecture.***

|  |  |  |
| --- | --- | --- |
| **Category** | **Parameter** | **Description/Examples** |
| Model architecture parameters | Model type | Transformer |
| input/output type | Eigenvector |
| training collaboration type | Type1 |
| latent message size | 57，104，270bit |

***Proposal 4. To consider standardizing the reference encoder and dataset for option 4.***  |
| [**R4-2409762**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409762.zip) | Samsung | *AI-CSI prediction:***Observation 1:** RAN4 agreement on CSI prediction accuracy metrics, i.e., throughput as default metrics for inference, is related to specifying RAN4 requirement only if the sub-use case of CSI prediction is confirmed for normative phase (decided in the checkpoints in RAN#105 (Sept ’24)).**Proposal 1:** For CSI prediction accuracy metric for inference, relative throughput (i.e., throughput by following predicted PMI over the baseline throughput) can be adopted, but* FFS the definition of baseline throughput (as denominator):
	+ Option 1-1: Throughput achieved by following random PMI (with the same codebook used for the reported predicted PMI)
	+ Option 1-2: Throughput achieved by following UE’s last reported PMI given by UE measurement

**Proposal 2:** For relative throughput used as CSI prediction accuracy metric for inference, Option 1-2 (Throughput achieved by following UE’s last reported PMI given by UE measurement) is adopted for baseline throughput. **Proposal 3:** No RAN4 discussion is needed on the testability issues for performance monitoring (including test method and test metrics to be used), until RAN1 clarify the details for monitoring.*AI-CSI compression:***Observation 2:** For TDL-based dataset for PMI reporting, the transformer-based encoder/decoder for AI-CSI compression is very easily to be over-fitted, with reasonable training dataset but much degraded validation dataset. **Observation 3:** The similarity performance on TDL-A validation dataset is slightly degraded if the TDL-A/C mixed dataset is used for training compared with only TDL-A dataset used. **Observation 4:** The over-fitting problem can be mitigated if the number of bits for latent message can be increased compared to the number used for Type-II codebook. **Proposal 4:** If the proponents of TDL-based channel model shall demonstrate the variance from TDL channel model is large enough to avoid an over-fitting model training with typical model design. **Observation 5:** 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 6:** 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.  |
| [**R4-2409782**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2409782.zip) | MediaTek inc. | **Proposal 1**: In RAN4 discussion, the reference encoder/decoder is used to derive the performance requirements.**Proposal 2**: RAN4 can consider the following two options for reference encoder/decoder:* Option 1: RAN4 discuss how to define reference encoder and decoder. RAN4 can only define some parameters, e.g., model type, layer size, and DUT vender can have some flexibility on implementation.
* Option 2: If RAN4 can conclude on test encoder, RAN4 can consider using the test encoder/decoder as the reference encoder/decoder.

**Proposal 3**: 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.
 |
| [**R4-2408659**](https://www.3gpp.org/ftp/TSG_RAN/WG4_Radio/TSGR4_111/Docs/R4-2408659.zip) | Nokia  | ***On alignment in two-side CSI use-case:*****Proposal 1: RAN4 to capture the AIML CSI compression scheme from Figure below for information to simplify the discussion of the simulation scenario/test case, model parameters, data sets, etc.**Figure 1: The scheme of AI/ML-based compressed CSI feedback performance evaluation**Proposal 2: RAN4 to split the study of two-sided CSI feedback at least in two stages:**1. **Task 1 – Parameters and performance alignment study**
2. **Taks 2 – Test/reference decoder/encoder derivation study**

**Proposal 3: RAN4 to consider the flow chart for Task 1 in Figure below to limit the set of model and training parameter and evaluate the feasibility of Option 3 and/Option 4.**Figure 2: Flow chart for Task1: Parameters and performance alignment study**Proposal 4: RAN4 to avoid using AI/ML models trained specifically for the test/simulation parameters, i.e., consider realistic and generic encoder and decoder training, e.g., based on CDL channel even if tested in TDL.****Observation 1:** Good alignment in accuracy metric A or performance metric M reported by different companies in Task 1 does guarantee interoperability in between the encoder and decoder implementations across the companies.**Observation 2:** Even if test decoder is fully specified (Option 3), it is not still obvious what encoder shall be assumed/can be used to derive the performance requirements and/or in the actual test.**Proposal 5: There is no need to defined reference decoder if test decoder is fully specified (Option 3).****Proposal 6: RAN4 to discuss which encoder can be used together with the fixed decoder (Option 3), considering at least the following options:****a. Option 1: (Reference) encoder is trained together with the test decoder and specified like test decoder.****b. Option 2: Encoder training based on test decoder is up to each company.****c. Option 3: Some parameters of (reference) encoder are specified, but its training is up to each company.****Proposal 7: RAN4 to consider the flow chart in Figure below as a possible way forward to the derivation of test decoder and performance requirements following Option 3.**Figure 3: A flow chart of possible process of derivation of test decoder and performance requirements with Option 3.***On Metrics/KPIs for AI/ML-based CSI feedback:*****Observation 3:** In the legacy requirements, relative thought was used to normalizes out implementation differences in precoding application. 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 8: RAN4 to adopt relative throughput metric γ based on random Type I PMI feedback for AI/ML-based CSI feedback performance requirements and use it in the comparison of simulation results.****Proposal 9: RAN4 to consider CSI prediction accuracy metric (e.g., SGCS) based on known CSI value for performance monitoring tests taking into account RAN1 specification of corresponding mechanism.*****On simulations, Model, and Training parameters:*****Proposal 10: RAN4 needs to consider three groups of parameters to evaluate and align the performance of AI/ML CSI compressions: 1) Performance evaluation/test case parameters 2) Model architecture parameters 3) Model training parameters.****Observation 4:** RAN1 LL simulation parameters are not detailed enough in comparison to RAN4 test parameters. RAN1 LL simulations parameters do not necessarily contradict the parameters based on RAN4 test cases.**Proposal 11: RAN4 configure test/simulation parameters based on one of the existing PMI reporting tests from TS 38.101-4 as a starting point, e.g., 6.3.3.1.4 Single PMI with 32TX TypeI-SinglePanel Codebook or 6.3.3.1.6 Multiple PMI with 16Tx Enhanced Type II Codebook as a starting point.****Proposal 12: RAN4 can also consider high-level LL simulation parameters aligned with RAN4 test cases from the table below.**High-level link-level simulations parameters aligned with RAN4 test cases

|  |  |
| --- | --- |
| Parameter | Value |
| Duplex, Waveform  | FDD OFDM  |
| Reference carrier frequency (for information) | 2GHz |
| Bandwidth | 40MHz |
| Subcarrier spacing | 30kHz  |
| Nt | 32: (8,4,2,1,1,4,4), (dH,dV) = (0.5, 0.8)λ and/or16: (8,4,2,1,1,2,4), (dH,dV) = (0.5, 0.8)λ |
| Nr | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ |
| Channel model | TDLA30-5 as starting point,CDL-C (3km UE speed, 30n delay spread) can be considered |
| Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline. |
| Rank per UE | Rank 1 |

**Proposal 13: RAN4 to consider the preferred values of parameters as described in Table below for the feasibility checking of Option -3 for AI/ML test decoder design.**Table: Proposed Option 3 Test Decoder Parameters [3]

|  |  |  |  |
| --- | --- | --- | --- |
| **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: 1Channel 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 options) | Parameters for Generalization Scenarios | UE speed, SINR, Indoor/outdoor, LOS/NLOS, Propagation model, etc. | FFS depending 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 options) | Scalability parameters | 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. |

**Proposal 14: We propose to incorporate training completion criteria into the parameters both Option 3 and Option 4. For Option 3, the completion criteria can be defined by the number of epochs, batch size, and learning rate, whereas for Option 4, the criteria can be specified as a minimum SGCS performance threshold.*****On Option 4 test decoder:*****Observation 5:** Data-set based approach (Freeze complete training data while leaving model architecture for implementation) is the best for Option 4 test decoder design to address interoperability.**Proposal 15: For Option 4-based test decoders, focus on the following sub-options, where Option 2a is currently preferred to best address interoperability:*** **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 relation to RAN1:*****Observation 6:** There is an inaccuracy in RAN1 conclusion about RAN1 Option 1: “It is RAN1’s understanding that Option 1 corresponds to RAN4 options, e.g., RAN4-Option3, or RAN4-Option4.” RAN1 Option 1 is defined as “Fully standardized reference model (structure + parameters)”. Hence, it corresponds only to RAN4-Option 3 and not to RAN4-Option 4.**Observation 7:** Even though RAN1 and RAN4 studies of the two-sided model design can continue rather independently before/if working itemed phase, an impact of RAN1 interoperability solutions on RAN4 test decoder design and performance is expected, especially based on the progress in RAN1 Option 1 and Option 3.**Proposal 16: RAN4 will need to evaluate the compliance of RAN4 reference/test encoder/decoder designs with RAN1 outcomes (e.g., whether RAN1 Option 1 reference model can be re-used for RAN4 Option 3, and whether RAN1 Option 3 model structure can be used for RAN4 Option 4).*****On Generalization:*****Observation 8:** 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 9:** 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.**Proposal 17: For the verification/testing of generalization/scalability related aspects in RAN4 for AI/ML enabled CSI feedback enhancement, RAN4 should define different scenarios based on parameters listed in the tables below.**

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

*Parameters for Generalization Scenarios*

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

*Parameters for Scalability Scenarios***Proposal 17: For the verification/testing of generalization/scalability related aspects in RAN4 for AI/ML enabled CSI feedback enhancement, RAN4 should define different scenarios based on parameters listed in the tables above.****Observation 10:** Less generalized functionalities across a set of scenarios can result in frequent switching of model/functionality resulting in performance degradation.**Proposal 18: RAN4 needs to design a new metric, indicative of generalization capabilities of AI/ML model/functionality, to verify the generalization performance of the model/functionality in different scenarios.*****On LCM Core requirements:*****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 19: 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.** |

## Open issues summary

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

1. Reference encoder/decoder definition
2. Option 3 split
3. RAN4 – RAN 1 alignment
4. standardization steps for Option 3
5. Reference encoder/decoder handling
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

*Reference encoder/decoder definition*

Definition of reference encoder/decoder was discussed in the previous meeting but was not agreed.

**Issue 4-1: Reference encoder/decoder**

* Proposals
	+ Option 1:
		- the encoder/decoder used in RAN4 discussions at least for simulation alignment/requirement derivation, test decoder derivation and/or test decoder verification. It could be documented (in TR, WF, etc) or captured in the specifications as necessary.Option 2: Other definitions
	+ Option 2:
		- Reference decoder/encoder: The decoder/encoder model used to define the minimum performance requirements. The Reference decoder is identical to the Test decoder.
	+ Option 3: other definition
* Recommended WF
	+ Option 1

To be discussed if any clarifications are needed

### Sub-topic 4-1bis

*CSI-compression scheme*

The scheme of AI/ML-based compressed CSI feedback for parameters and metric alignment.

**Issue 4-1bis: AI/ML based CSI compression scheme**

* Proposals
	+ Option 1: Agree on the baseline scheme for CSI compression use-case.
	+ Option 2: other
* Recommended WF

To be discussed if any clarifications are needed.

### Sub-topic 4-2

*Standardization steps for Option 3*

A flow chart for the work on Option 3 was presented in the previous meeting (R4-2405653, reproduced below) and discussed. This was discussed, however there was no formal agreement. A refinement to this chart was proposed in R4-2407236. Also, some proposals were made in R4-2407368. These should be discussed to further the align the steps needed to progress on Option 3 feasibility.

**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: Flowchart below better represents the steps needed in RAN4, should be considered as baseline:

****

* + Option 2: Following changes should be made to the chart from R4-2405653(discussed in RAN4#110Bis)
		- Add a step, “identify target cases with specific test conditions” before step 1.
		- For step 1, suggest to limit the scope of potential model types. Transformer-based could be prioritized.
		- Add a step, “Define evaluation methodology” before step 4.
		- Suggest to align the understanding of step 5 on, whether only a single model architecture would be determined for 2-sided case at least for this release.
		- Discuss the necessity of step 8 and 9.
	+ Option 3: other changes
* Recommended WF

To be discussed

### Sub-topic 4-3

*Decoder parameters for Option 3*

Several companies made proposals for information that should be agreed in order to be able to derive a full decoder for Option 3. Some agreements on needed parameters were reached in previous meetings and documented.

 **Issue 4-3: Decoder parameters**

* Proposals

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 　 | Vivo (updated) | CATT | QC (updated) | Xiaomi | Intel | E/// (updated) | Apple (updated) | Nokia (updated) | Keysight | ZTE (updated) | Samsung (updated) |
| Model type | Transformer, CNN, RNN, MLP | Transformer | 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 | Transformer, CNN, RNN, MLP | Transformer | 　 | Transformer | Transformer |
| Model depth | Number of layers | 　 | 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]) | 　 | 　 | 4 layer |
| Layer type | Fully connected, convolutional, activation layer, normalization layers, etc. | 　 | 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. | 　 | 　 | Fully connected layers with activation function for each attention layer/block |
| Layer size | Neuron count and configuration | 　 | 　 | 　 | 　 | 　 | Neuron count and configuration | *Note that output layer can be different.* | 　 | 　 | Scalar quantizer, 2 bits per dimension |
| Quantization method for the encoder output | Scalar, vector (with codebook) | 　 | 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. | 　 | 　 | 142bits (for 2-layer case) |
| Encoder-decoder interface | Number of bits of latent message | 　 | Use power of 2, choose from 32, 64, 128, 256 and 512 bits. | 60, 120, 280bits | 　 | Consider 63, 110 or 230 | Number of bits of latent message | FFS, e.g., 64 latent dimensions with two-bit quantization, i.e., 128 overhead bits. | 　 | 57，104，270bit | 86bit (for 1-layer case) |
| Fixed point representation | Int8, int16, floating point etc | 　 | 　 | 　 | 　 | Int8, int16, floating point etc. | Int8, int16, floating point etc | FFS, decision to be made during/after model design, or may be left for implementation. | 　 | 　 | N/A |
| Format of input to encoder/output of decoder | 　 | 　 | 　 | 　 | 　 | Consider pre-processing of Eigenvector using Enhanced Type 2 codebook | 　 | Eigen vectors,Sub-band reporting (e.g., [13] sub-bands for 10 MHz CBW, 15kHz SCS). | 　 | Eigen vectors | Eigenvector |
| 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. | 　 | 　 | 　 | 　 | 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, etc. | 　 | SGCS | 　 | 　 | NMSE | 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. | 　 | 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,  | 　 | 　 | Learning rate, batch size, regularization techniques and strength, optimization algorithm, etc. | 　 | 　 | 　 | 　 |
| 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: which parameters are needed and what should the values be

### Sub-topic 4-4

*Simulation parameters for Option 3*

Several companies brought proposals on simulation parameters to be used for encoder/decoder derivation to further progress the feasibility of option 3. These should be discussed and a set of parameters should be agreed.

**Issue 4-4: Simulation parameters for Option 3**

* Proposals are shown in the table below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 　 | CATT | Nokia | Intel | Vivo | QC |
| Duplex, Waveform  | FDD, OFDM  | FDD OFDM | FDD  | FDD OFDM  | FDD OFDM  |
| Carrier frequency | 2GHz | 2GHz | 　 | 2GHz  | 　 |
| Bandwidth | 10MHz | 40MHz | 10 MHz | 20MHz | 20MHz |
| Subcarrier spacing | 15kHz for 2GHz | 30kHz | 15 kHz  | 15kHz  | 15kHz  |
| Nt | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ | 32: (8,4,2,1,1,4,4), (dH,dV) = (0.5, 0.8)λ and/or 16: (8,4,2,1,1,2,4), (dH,dV) = (0.5, 0.8)λ | CDL channel models- 32 ports: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ- 16 ports: (8,4,2,1,1,2,4), (dH,dV) = (0.5, 0.8)λTDL channel models- Low correlation | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ | 32: (8,8,2,1,1,2,8), (dH,dV) = (0.5, 0.8)λ |
| Nr | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ | CDL channel models- 4RX UE: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ for (rank 1-4)TDL channel models- 4RX UE, Low correlation | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ | 4: (1,2,2,1,1,1,2), (dH,dV) = (0.5, 0.5)λ |
| Channel model (Delay spread) | CDL-C (300 ns) as baseline, CDL-A (30 ns) as optional. | TDLA30-5 as starting point, CDL-C (3km UE speed, 30n delay spread) can be considered | 1) CDL-C 300ns DS, 100Hz (TR 38.901), 2) TDLC300-100 (TS 38.101-4) | CDL-C | CDL-C or TDL-A |
| UE speed | 3kmhr | 　 | 　 | 3kmhr | 10Hz doppler |
| Channel estimation | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline.Ideal DL channel estimation is optionally taken into the baseline of evaluation methodology for the purpose of calibration and/or comparing intermediate results (e.g., accuracy of AI/ML output CSI, etc.). Up to companies to report whether/how ideal channel is used in the dataset construction and performance evaluation/inference. | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline. | Realistic channel estimation algorithms MMSE-IRC as the baseline receiver | Realistic channel estimation algorithms (e.g., LS or MMSE) as a baseline.Ideal DL channel estimation is optionally taken into the baseline of evaluation methodology for the purpose of calibration and/or comparing intermediate results (e.g., accuracy of AI/ML output CSI, etc.). Up to companies to report whether/how ideal channel is used in the dataset construction and performance evaluation/inference. | Realistic channel estimation algorithms (e.g., LS or MMSE) or ideal DL channel estimation |
| Rank per UE | Rank 1.  | Rank 1 | Fixed rank 1 as the starting point4 PRB subband size | Rank 1-4. Companies are encouraged to report the Rank number, and whether/how rank adaptation is applied | Rank 1 |
| Evaluation metric(s) | SGCS | 　 | SGCS metric (as described in TR 38.843 6.2.1) or NMSE | 　 | 　 |
| Delay spread | 　 | 　 | 　 | 30ns | 30ns  |
| Latent message size | 　 | 　 | 　 | 　 | Use power of 2, choose from 32, 64, 128, 256 and 512 bits. |

* Recommended WF
	+ Parameters to be discussed and agreed

### Sub-topic 4-5

*Option 4 for 2-sided model*

Several companies brought proposal on how to further study/check the feasibility of option 4.

**Issue 4-6: Option 4 for 2-sided model**

* Proposals
	+ Option 1: Qualcomm (R4-2407334)
* **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/decoder pair**
	+ **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/decoder as a reference encoder/decoder pair 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 2: Xiaomi (R4-2407847): ***Standardized data / dataset format + Dataset exchange between NW-side and UE-side***
* Option 3: vivo(R4-2408294)
	+ **Option 4a: Reference encoder + test decoder model structure, and channel generation method are in the spec;**
	+ **Option 4b: Test decoder model structure + reference encoder model structure, and dataset (PMI and corresponding channel) are in the spec;**
	+ Option 4 – Huawei (R4-2409003))
* 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: Nokia (R4-2408659)
* **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 6: Ericsson (R4-2408492)
		- the latent space needs to be standardized
		- standardize a reference encoder in order to capture the latent space
	+ Option 7: 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 4-6

*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 4-6: 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