**3GPP TSG-RAN WG4 Meeting #114bis R4-2504685**

**Wuhan, China, 7th ‒ 11th April, 2025**

**Agenda item:** 7.20.3

**Source:** Moderator (Ericsson)

**Title:** Topic summary for [114bis][127] NR\_AIML\_air\_part2

**Document for:** Information

# Introduction

This topic summary covers two areas:

* General topics, considering both the WI and the SI
* Two-sided CSI compression

For the general topics, discussion is broken down into LCM related topics (in particular delay requirements), generalization, post-deployment handling and other topics.

For CSI compression, an attempt is made to capture some observations from the simulation results. Although more results are needed and the observations will be updated, the hope is that discussing what can be learned from the results presented so far can help in focussing thought on how to progress with the simulations and where the results so far are leading. Complementing this moderator summary is an Excel sheet capturing the results, which can help in reviewing the potential observations and in providing insight into where further simulation effort and results are needed.

There are also a number of issues that do not directly relate to simulations, such as what is to be captured in the specification for each of the options for 2-sided CSI compression.

# Topic #1: General issues for the WI and SI

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| R4-2503327 | Korea test laboratory | **Observation 1: “Post deployment management based on LCM” requires frequent signaling exchanges between UEs and the network.**  **Proposal 1: Optimize LCM-based performance monitoring in post-deployment testing to reduce unnecessary signaling.**  **Proposal 2: Define Fine-tuning as the process of adjusting existing model weights while retaining the pre-trained model’s structure. Fine-Tuning modifies only a subset of model parameters, allowing for targeted adaptation to new data without altering the overall model architecture. Due to the minimal impact on structural integrity, conformance testing requirements for fine-tuned models should be limited to the functionality configuration.**  **Proposal 3: Define Model Update as a process that involves full replacement of an AI/ML model, requiring retraining from scratch. Model updates may also include architectural modifications, such as transitioning between different deep learning frameworks (e.g., CNN to Transformer). Due to the substantial changes in model behavior, conformance testing must be conducted comprehensively, covering both functional validation and real-world performance testing.**  **Proposal 4: The difference metric shall be used for performance monitoring in post-deployment testing and the metric used for monitoring in normal operation.**  **Observation 5: Current agreements (RAN1 #116-bis, Step C/D) do not explicitly mandate signaling of UE-side model updates, creating ambiguity in certain model ID assignment scenarios.**  **Proposal 5: Clearly define and introduce an explicit UE signaling procedure (AI-ML-Model-Update-Report) within the core 3GPP specifications, which becomes mandatory if post-deployment validation relies solely on network-side validation (NW-sided only).**  **Observation 6: It has different behavior to verify the minimum performance of AI/ML functionality/feature** **depending on “Signalling procedures for model and functionality LCM” in TR 38.843.**  **Proposal 6: It should be discussed within the RAN4 group regarding the efficient process for conformance testing and post-deployment testing, using fallback mechanisms or deactivation, especially if it might not be possible to define the mechanism in detail at RAN1/RAN2 since the reason of the implementation area.**  **Observation 7: Although UE-side field-data-based training inherently mitigates overfitting risks, the recently agreed RAN1 procedure (AI-Example1) further emphasizes the need for explicit post-deployment validation mechanisms to confirm generalization performance.**  **Proposal 8: It is proposed that RAN4 explicitly incorporates overfitting validation mechanisms into post-deployment scenarios. Such mechanisms should define clear procedures for utilizing independent validation datasets, performance monitoring, and corresponding NW-UE signaling to robustly detect and mitigate overfitting risks in field-deployed AI/ML models.** |
| R4-2503229 | CAICT | **Proposal 1: Prioritize option 2 (performance monitoring and LCM) for post-deployment handling.**  **Proposal 2: For LCM monitoring, suggest RAN4 to consider specifying requirements on delay requirements of monitoring reporting and model switching/fallback.**  ***Observation 1:*** *For beam management, RAN1 agreed to support Type-1 performance monitoring, which includes case 1 (NW-decision and NW-initiated) and 3 (UE-decision, reporting and NW-initiated).*  ***Observation 2:*** *For positioning case 1, RAN1 agreed to support option A of performance monitoring, which includes case 2 (NW-decision and UE-initiated) and 3.*  **Proposal 3: Following cases could be considered when discussing delay requirement of AI fallback to non-AI, including:**   1. **NW-decision and NW-initiated** 2. **NW-decision and UE-initiated** 3. **UE-decision, reporting and NW-initiated** |
| R4-2503523 | Qualcomm | **Observation 1: Legacy non-AI-ML CSI measurement requirements were discussed in RAN1 and entirely captured in RAN1 spec.**  **Observation 2: Two of three UE sided AI-ML positioning use cases (case 2a/2b) have been removed from the AI-ML air-interface WID during the last RAN plenary.**  **Observation 3: The starting time to activate periodic AI-ML based beam management resources are not still finalized in RAN1/RAN2.**  **Observation 4: The end point of deactivating AI-ML based beam management is not clear. The steps to fall back to non-AI-ML beam management after deactivation requires additional configurations and transmissions of network (e.g., configuration of setA resources).**  **Proposal 1: RAN4 should not define AI-ML based CSI activation/switching delay requirements and let RAN1/RAN2 to define it.**  **Proposal 2: RAN4 should wait to see the progress of case 1 before discussing activating and switching requirements of UE sided AI-ML based positioning.**  **Proposal 3: RAN4 should wait until the starting time of periodic and semi-persistent setB resources for AI-ML based beam management are finalized in RAN1.**  **Proposal 4: RAN4 considers defining the delay requirements to activate a functionality in AI-ML beam management where setB resources are aperiodic.**   * **Note: The timeline starts when UE receives the network command that triggers the of setB resources. The timeline ends when UE is expected to convey the inferred properties of setA resources via uplink.**   **Proposal 5: RAN4 does not define deactivation delay requirements for AI-ML based beam management.** |
| R4-2503306 | Xiaomi | **Proposal 1: For beam ID prediction accuracy, there is no performance requirement in legacy. Set a specific requirement for AI.**  **Proposal 2: For RSRP prediction accuracy, measurement error may have impact on whether AI requirements can exceed “legacy” requirements.**  **Observation 1: It’s possible that similar KPI, i.e. Top K/M, is defined for conformance test and performance monitoring for beam ID prediction accuracy.**  **Proposal 3: Suggest to design the similar metric for performance monitoring and conformance test. Further discuss the condition to skip option 1.** |
| R4-2503514 | Apple | ***RAN4 performance testing goals***  **Proposal 1: Deprioritize Option 1 and focus on Option 2 for performance requirements as a testing goal**  **Proposal 2: For verifying performance gain of AI/ML models/functionalities, RAN4 can define multiple independent** **test cases with different scenarios and configurations/conditions as reported through UE capability signaling, which could include:**  **Propagation conditions/scenarios, e.g., channel modes defined for different scenarios (CDL, AWGN, etc.) in TR38.901, Doppler conditions, SNR levels, various deployment scenarios (Uma, Umi, ISD), different cell with different gNB height, various indoor/outdoor UE distributions, various UE mobility etc.**  **- Configurations:**   * **Various UE parameters: number of UE beams, antenna panels, antenna array dimensions, different AI/ML configurations: different set B of beams, T1 for measurement/T2 for prediction for BM-case2** * **Various gNB settings: Tx codebooks (set A/B beams), beam widths, antenna spacings, Tx antenna dimension, antenna port layouts, TXRU antenna virtualization (e.g. (8,8,2,1,1,2,8), vs (8,4,2,1,1,2,4))**   **RAN4 should conduct a down-selection process to narrow down the scenarios and configurations (parameters) for the conformance test. This test will serve as a benchmark for evaluating the model's performance across a range of unseen scenarios and configurations.**  ***Static/non-static scenarios/configurations***  **Proposal 3: If non-static scenarios/configurations are supported for certain use cases, they can be included as part of generalization tests.**  **Proposal 4:** **Non-static scenarios/configurations should be considered for test cases only if static scenario/configuration testing fails to fulfill the testing objectives. CSI and beam management temporal prediction use cases are particularly suitable for introducing non-static environments during testing**  ***Generalization/scalability aspects***  **Proposal 5: To determine the granularity of additional scenarios/conditions for defining generalization tests for each use case, it's essential to study how the AI model's behavior changes with different scenarios and conditions.**  **Proposal 6: To improve the generalization behavior of the model, training with a diverse dataset should be investigated.**  **Proposal 7: Investigate the reduction of generalization tests by training with a mixed dataset containing samples from different configurations. Investigate the definition of a single generalization test where the testing data consist of a random mixture of these configurations. If the performance degradation between the model trained on the mixed dataset and tested with random mixture configurations and the model trained and tested specifically for each configuration meets a predefined performance margin criterion, the test could pass.**  Proposal 8: Study post-deployment procedures to augment conformance testing for effectively managing performance across all possible deployment conditions/scenarios (which are not tested)  **Proposal 9: For UE-side models and/or UE-part of two-sided models** **it is suggested that the scenarios and configurations utilized for generalization tests can be determined based on the supported configuration reported by the UE as part of capability signaling.**  **Proposal 10: In the RAN4 core requirement, it is mandated that the consistency or association between of additional conditions during both training and inference is guaranteed. This serves as additional assistance information from the network side for testing UE-sided models during inference and monitoring**  **Proposal 11: RAN4 will explore methods to specify the identified scenarios and/or configurations per use case in future release, contingent upon other WGs can specify the granularity and the capability signaling.**  **Proposal 12: For defining generalization tests, RAN4 should define identified scenarios associated with the UE capability report of an AI/ML-enabled Feature FG, and other scenarios (additional conditions). RAN4 should also define minimum level of performance for the identified scenarios and/or conditions.**  **Proposal 13: Other scenarios and/or configurations can be interpreted as the scenarios and/or configurations that are not reported by UE capability signaling for an AI/ML-specific functionality or model ID.**  *Scenario/configuration specific Models (Fine-tuning)*  **Proposal 14: RAN4 to discuss the practicality of formulating a framework that facilitates on-device fine-tuning. The focus will be on exploring the feasibility of creating a dynamic and site-specific approach to online training and fine-tuning (e.g reinforcement learning)**  **Proposal 15: For the purpose of limiting the amount of data needed for fine-tuning and reducing the number of fine-tuning iterations investigate transfer learning and meta-learning for efficient re-learning across different scenarios to aid generalization**  **Proposal 16: UE to update its stored AI/ML models with the new model ID after fine-tuning, where this ID can be associated with the training data (which implicitly have the additional conditions) used to fine tune the model.**  Proposal 17: RAN4 should investigate the options for enhancing the generalizability of AI/ML models by providing the appropriate assistance/side information as input signal to the inference engine of the AI/ML model and discuss the feasibility of training with diverse datasets across different additional conditions Principles on the definition of requirements Proposal 18: RAN4 should clarify/agree that the side conditions of the testing procedures should remain the same for legacy and AI/ML methods. Requirements for LCM (Performance Monitoring) **Proposal 19: RAN4 shall define RAN4 core requirement for performance monitoring tests based on RAN1/2 defined monitoring metrics/methods for particular (sub-)use case**  **Proposal 20: RAN4 shall consider the latency requirements for model monitoring input data as well as the establishment of tolerance margin requirements for the specified KPIs for model monitoring per use case**  **Proposal 21: RAN4 to investigate a performance monitoring framework to detect issues such as data drift, model degradation, and AI/ML model aging, triggering actions like re-training, re-configuration, or model switching. The system will ensure high detection accuracy, real-time monitoring, and proactive predictions, enabling reliable adaptation to changing channel conditions while maintaining optimal performance.** Post Deployment validation and On-Device Fine tuning **We are making a new proposal for post deployment (Option 3) combined with on-device fine tuning. Based on the principles below:**  **Proposal 22: 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 23: 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 24: 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 25: The UE updates its model list with a new model derived from either fine-tuning, model transfer from the NW, or monitoring to ensure the consistency of additional conditions. Then, the UE assigns an ID to the new model that supports the NW's additional conditions and shares this information with the NW. If some of the new conditions/configurations are standardized, the UE updates its capability report accordingly.**  **Proposal 26: If there is no ID available to associate training data with additional conditions, and monitoring procedure fails to guarantee the consistency of the model with the additional conditions then UE should fallback to legacy mode.**  **Proposal 27: *(Proposed Option 3 for post deployment)* For the post deployment framework, investigate an on-device or server-side model monitoring and performance assessment framework to track both active and inactive AI/ML models that experience actual UE hardware and field data (could be uploaded to the server). The aim is to dynamically manage the model database, enabling proactive monitoring for computing and updating key performance indicators (KPIs). This will facilitate seamless transitions to newly updated and fine-tuned models as needed.**  **Option 3 captures: Option 2 (LCM) + Model monitoring for both active and inactive models + KPI tracking for active/inactive models + model update/finetuning with few samples learning**  **Additionally, option 3 captures concepts from Option 2 (LCM) along with (A) model monitoring for both active and inactive models, (B) KPI tracking for active/inactive models and (C) model update/finetuning with few samples learning:**   * **Validation takes into account the UE hardware in which the model is to be deployed/activated** * **Validation takes place in the field (i.e., on each individual device) or in the server with data sent from the device (thus capturing UE HW and actual field data)** * **Validation can be done at the device (inactive models) along with inference for another active functionality(ies) (active and inactive models)** * **Using performance monitoring and LCM procedures** * **Monitoring can be used for managing fallback, change in functionality, model update/model switching/model transfer, if applicable**   **To minimize the number of samples required for fine-tuning/model update and reduce training costs, the framework will leverage Transfer Learning and Meta Learning techniques, ensuring efficient model management. The ultimate goal is to optimize performance, reduce overhead, and ensure models adapt seamlessly to changing conditions for post deployment and fine-tuning in a proactive fashion.**   * + **KPIs based on monitoring procedures are used to update the monitoring scores for all models at the UE side (active or inactive models stored). Model with currently best KPI (that meets requirements) is deployed**   + **KPIs computed at UE can be signaled to the NW, KPIs are used to update the monitoring scores for each model (updates the database of active and inactive models). NW can decide if any of the models need update or fine tuning. It requests a small amount of data from the UE for fine tuning (data captures UE HW and field data). NW fine tunes the model and sends only the updated parameters along with the model ID. Based on KPIs the NW can indicate the model to be deployed or switched to at UE side or instruct the UE to fallback into legacy.**   + **NW can predict performance degradation (in future) and signal UE to a) switch model or (b) to identify which model to fine-tune**   **Proposal 28 *(Option 3)* : RAN4 to employ the post-deployment procedures described through the dynamic management of an AI/ML model database and RRC signaling for enabling monitoring procedures** Proposal 29 *(Option 3)*: In the post-deployment framework, models would be allowed to undergo fine-tuning, utilizing the model monitoring dataset for ongoing fine-tuning and re-training purposes. **Proposal 30: In the post-deployment framework, the network (NW) can indicate which model should be fine-tuned, with the fine-tuning process occurring either at the UE or within the network itself or an OTA server**  **Proposal 31 *(Option 3)*: If fine-tuning takes place at the network (NW) or an over-the-air (OTA) server, the UE transmits a small sample dataset with high precision (e.g., PMI) over the air. The NW or OTA server (where inference takes place) then performs fine-tuning using transfer learning or meta learning techniques on the identified model to be fine-tuned, minimizing the number of parameters that need updating. It subsequently requests another dataset for validation to validate the model for post deployment. The updated set of parameters is then downloaded to the UE, with the option to quantize the coefficients to reduce overhead.**  **Proposal 32 *(Option 3)*: If fine-tuning occurs at the UE, the UE is responsible for performing the fine-tuning using a small sample training dataset, with the network (NW) signaling which model to undergo updating. The UE then performs model monitoring to verify and validate the model post-deployment. The goal is to achieve optimal model performance as quickly as possible, minimizing the number of samples required during Transfer Learning or Meta Learning to reduce training costs and achieve fast adaptation.** 2.7 Data collection for testing **Proposal 33: 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.** |
| R4-2503770 | CMCC | ***Proposal 1: the switch delay from non-AI to AI is different for different report type:***   * ***for periodic CSI reporting, since UE can autonomously activate the applicable functionalities, the starting point of switch delay from non-AI to AI is that UE report applicable functionalities via RRCReconfigurationComplete*** * ***for semi-persistent and aperiodic CSI reporting, the starting point of the switch delay from non-AI to AI is the point receiving activation indication by MAC CE/DCI***   ***Proposal 2: for CSI prediction, it is proposed to use intermediate KPI, e.g. SGCS or NMSE, as requirements/tests metrics for LCM.***  ***Proposal 3: for AI/ML based beam management, AI/ML based performance requirement should not be worse than the legacy measurement performance.*** |
| R4-2503920 | Ericsson | * ***Proposal 1 (pre-activation functionality/model update testing)****: Post deployment validation should be at the facility that creates the update.* * ***Proposal 2 (pre-activation functionality/model update testing)****: To move the discussion on post-deployment validation forward, ask RAN5 to clarify the GCF procedure and whether anything similar might be considered.* * ***Proposal 3 (pre-activation functionality/model update testing)****: Further discuss “major” and “minor” model changes.* * ***Observation 1 (post-deployment management based on LCM)****: It is not obvious as yet whether monitoring can useful identify individual model behavior and whether it can be interoperable.* * ***Proposal 4 (post-deployment management based on LCM)****: Continue to work on monitoring to understand how useful and interoperable it can be before deciding on whether it could be a sole solution for post-deployment performance assurance.* * ***Proposal 5 (LCM requirements)****: The following definition for AI/ML functionality switching can be assumed by RAN4 in its AI/ML discussions:*   ***AI/ML functionality switching*** *is a situation when one AI/ML functionality is deactivated and another functionality is activated by the same message.*   * ***Proposal 6 (LCM requirements)****: The following definition for a fallback from AI/ML operation to non-AI/ML operation can be assumed by RAN4 in its AI/ML discussions:*   ***Fallback*** *from AI/ML operation to non-AI/ML operation is a situation when one or more AI/ML functionalities are deactivated and there is no relevant activated AI/ML functionality left.*   * ***Proposal 7 (LCM requirements)****: No fallback delay requirement for AI/ML positioning use case 1 is necessary, since AI/ML positioning use case 1 is configured by LMF as a separate positioning method, based on the current RAN2 agreements. (The statement can be rediscussed if there will be new related RAN2 agreements impacting the procedure.)* * ***Proposal 8 (LCM requirements)****: RAN4 will consider defining functionality activation delay requirement for the beam management use case at least with periodic CSI reporting, FFS for aperiodic and semi-persistent reporting.* * ***Proposal 9 (LCM requirements)****: RAN4 will consider defining functionality deactivation delay requirement for the beam management use case.* * ***Proposal 10 (LCM requirements)****: FFS: whether fallback delay includes other delays than AI/ML functionality deactivation delay for beam management (e.g., a delay until the next non-AI/ML report).* * ***Proposal 11 (LCM requirements)****: A requirement for the maximum delay until the UE actually starts the data collection after receiving the network configuration can be considered for data collection, to ensure that the data collection is not delayed. Alternatively, a UE behaviour can be specified that the UE shall start, e.g., from a first signal occasion available after receiving and decoding the command.* |
| R4-2504229 | Nokia | **Proposal 1: The RAN4 discussions to use “post-deployment testing” defined as “the testing of functionality/configuration in a UE which is operating in the field, connected to a mobile network”.**  **Proposal 2: The RAN4 discussions to assume Option 1 described as “Post-deployment pre-activation functionality/configuration update testing”.**  **Proposal 3: The RAN4 discussions on the functionality/configuration which requires testing before it can be RRC/MAC activated by the NW for inference, to use** **“pre-activation” term to indicate “the functionality/configuration is available in (and supported by) the UE device, and it is not RRC/MAC activated”.**  **Proposal 4: The RAN4 discussions to assume Option 2 described as “Post-deployment post-activation functionality testing based on performance monitoring”**  **Proposal 5: The RAN4 discussions to use “test-active” defined as “the functionality/ configuration which is available in (supported by the) UE, it is RRC configured by the gNB/NW for post-deployment testing and, has been RRC/MAC activated by the gNB/NW for post-deployment testing”**  **Proposal 6: RAN4 should discuss on the framework related to the test-active mode of the ML-enabled feature to realize the post-deployment testing in Option 1.**  **Proposal 7: RAN4 should further discuss the possible in-field configurable conditions and requirements for a post-deployment test of an ML-enabled feature in Option 1.**  **Observation 1:** RAN4 to consider the need and feasibility of requirements and tests to ensure consistency and accuracy of monitoring metrics or other monitoring related data, for the configured, active and applicable functionality/ configuration, sent from the UE, and set requirements as needed.  **Proposal 8: RAN4 to consider the need and feasibility of requirements and tests to ensure the required functioning of performance monitoring procedures for the test-active functionality/ configuration, and set requirements as needed**  **Proposal 9: Based on the current RAN1- RAN2 agreements for Release 19, RAN4 to not purse testing procedures for individual UE-side ML model performance testing purposes.**  **Proposal 10: Based on the current RAN1- RAN2 agreements for Release 19 on performance monitoring, RAN4 to study the necessity and feasibility of testing requirements to ensure consistent performance monitoring between different UEs can be achieved.**  **Observation 2:** Frequent updates to the UE specific ML models supporting the ML-enabled feature configurations may significantly prolong the turnaround time and testing overhead if tested the lab.  **Proposal 11: Conformance testing for changed/updated ML functionality before deployment (Option 1) in the lab can be time consuming and it must be de-prioritized.**  **Proposal 12: RAN4 to ensure and to clarify that at least one version of the ML-enabled feature configuration (including the supporting ML models), that has passed conformance testing shall always be available in the device.**  **Observation 3:** In order to validate the new ML-enabled feature configuration before using it in the field, the UE need to support inference either in consecutive/time splitting manner or in parallel.  **Proposal 13: RAN4 to further discuss the option of using both active as well as test-active in parallel to realize post-deployment testing of updated ML-enabled configuration in the field.**  **Observation 4:** It should be noted that the Option 1 is not fully supported by RAN2.  **Proposal 14: RAN4 should further discuss on the potential signaling requirements to realize the post-deployment tests in Option 1.**  **Observation 5:** In order to realize post-deployment testing in Option 2, the signaling requirements can be part of the UE-side “LCM” signaling that are being discussed in RAN2.  **Proposal 15: Updates to current UE-side “LCM” signaling procedures are required to realize post-deployment testing of updated ML-enabled functionality in Option 2 as a part of performance monitoring. This must be done in alignment with RAN2 and impacts to RAN4 should be further discussed.**  **Proposal 16: Relation to GCF should be discussed in RAN5 when RAN5 discussions start.**  **Proposal 17: RAN4 to consider a combination of Option 1 and Option 2 as the best way forward solution for the post-deployment testing.**  **Proposal 18: In case of NW sided performance monitoring, UE should follow legacy measurement period requirements for the measurement of monitoring RS resources.**  **Proposal 19: In case of UE-assisted performance monitoring, RAN4 should consider defining delay requirement for performance metric reporting.**  **Proposal 20: Tolerance shouldn’t be defined as the accuracy of the performance monitoring report.**  **Proposal 21: Tolerance for performance monitoring should be defined as the level of performance degradation that needs to be reached before taking any LCM related action. RAN4 should not define any requirement for tolerance level for both NW and UE sided LCM.**  **Proposal 22: RAN4 to adopt that, according to RAN2,** **AI/ML functionality activation refers to the process of enabling an applicable functionality to perform inference.**  **Observation 6:** According to RAN2, one example of cases where the network deactivates a functionality is when it becomes non-applicable.  **Proposal 23: RAN4 to adopt that, according to RAN2, AI/ML functionality deactivation refers to the process of network deactivating an active functionality.**  **Proposal 24: For periodic CSI reporting in UE-sided beam management, the starting point of the AI/ML functionality activation delay is when UE receives RRCReconfiguration message with the configuration of the AI/ML functionality.**  **Proposal 25: For aperiodic and semi-persistent CSI reporting in UE-sided beam management, the starting point of the AI/ML functionality activation delay is when UE receives the MAC CE/DCI command to activate the functionality.**  **Proposal 26:** **The ending point of the AI/ML functionality activation delay is when UE starts reporting inference to the network.**  **Proposal 27: The starting point of the AI/ML functionality deactivation delay is when UE receives the command to deactivate the functionality from the network.**  **Proposal 28: RAN4 to monitor RAN1/RAN2 agreements about AI/ML functionality deactivation to revisit the starting point definition and define the ending point.**  **Observation 7:** Specifying different LCM procedures for the different one-sided use cases implies the need for defining use case-specific LCM-related delay requirements (e.g., activation/deactivation delays)  **Proposal 29: RAN4 to define use case-specific LCM-related delay requirements following the LCM procedures defined by RAN1/RAN2**  **Observation 8:** The way how test configurations/parameters and requirements in RAN4 will be defined will depend on the way how Associated IDs will be standardized in RAN1 and RAN2.  **Proposal 30: Each specified AI/ML-enabled Feature shall have RAN4 requirements and UE needs to pass at least one test for specific sub-set of applicable Feature configurations (matching functionalities) as part of conformance testing.**  **Proposal 31: A minimum set of test configurations (matching Associated ID(s)) shall be defined as mandatory for testing of AI/ML-enabled Feature.**  **Observation 9:** A considerable change of NW-side conditions that require the change of Associated ID may require RRC reconfiguration for the UE.  **Proposal 32: RAN4 does not need to define performance requirements that imply the change of NW configuration/Associated ID during the test.**  **Observation 10:** The number of additional requirements per given UE AI/ML-enabled Feature and Associated ID will depend on their concrete definitions of the functionalities/configurations in the other WGs and the generalization capabilities of the UE implementation solutions, and both require further study in per use-case manner.  **Proposal 33: For the testing of Generalization aspects, RAN4 needs to study the variation in the AI/ML-enabled Feature performance due to the change in test conditions (e.g., propagation conditions) to conclude about the final number of requirements/tests per given AI/ML-enabled Feature and Associated ID combination.**  **Proposal 34: The static radio scenarios/conditions for a given UE AI/ML-enabled Feature and/or Associated ID shall be assumed as baseline for RAN4 requirements purposes.** |
| R4-2504316 | Huawei, HiSilicon | ***Proposal 1:*** UE initiated transition may not be testable if UE-side proprietary information is involved, e.g., whether the AI model is available at UE anymore.  ***Proposal 2:*** RRM requirement of transition delay from AI to non-AI may be needed for NW initiated transition cases.  ***Proposal 3:*** RAN4 to discuss RRM requirement of transition delay with different definitions of non-AI.   * Option 1: AI method is deactivated only. * Option 2: AI method is deactivated and non-AI based method is switched to.   ***Proposal 4:*** If the AI related configurations/parameters is a subset of (or the same as) the legacy test configurations/parameters, take the existing test configurations/parameters as a starting point.  ***Observation 1:*** For some cases, to verify AI performance, new test set up is under discussion.  ***Proposal 5:*** If new test set up is introduced for AI test, two options are identified for defining AI performance requirement.   * Option 1: Define non-AI requirements under new test set up subject to AI. AI requirement has to be higher than that of non-AI.   + FFS: Feasibility of aligning non-AI performance baseline under new test set up. * Option 2: Define absolute requirement of AI without considering the comparison with non-AI.   + FFS: How to verify the AI benefits. |
| R4-2502409 | CATT | **Observation 1: For changes due to update or fine tuning of models,**   * **Validation in lab conditions cannot ensure a consistent performance between field and post deployment testing, because the updates or fine tuning are scenario-dependent.** * **Validation in the field is the same as the performance monitoring procedure.**   **Observation 2: For changes due to addition/removal of models, applicability checking procedure of functionality can handle the potential impacts on functionality performance.**  **Observation 3: For impacts due to data drift/mismatch, the post deployment testing is actually trying to identify the reasons why AI/ML functionality performs badly in the field, which is beyond the scope of RAN4 performance testing and belongs to management of AI/ML functionality.**  **Proposal 1: Impacts due to changes of models and data drift / mismatch be handled by at least the following procedures:**   * **Performance monitoring and LCM procedures** * **applicability checking procedure of functionality**   **Proposal 2: RAN4 discuss Option 1 after proponents clarify their motivations and its feasibility is proved.**  **Observation 4: Number of requirement set(s) depends on the granularity of AI/ML functionality which may refer to UE capability.**  **Proposal 3: RAN4 postpone discussing the number of requirement set(s) after RAN1 achieve agreements on UE capability.**  **Proposal 4: RAN4 to discuss the number of requirements per requirement set in performance phase, at which time RAN4 will have concrete information about UE capability and measurement configurations.** |
| R4-2503519 | Apple | **Observation 1: Input CSI data distribution can be used to detect a monitoring event. By detecting monitoring events without ground truth CSI excessive overhead signaling can be avoided**  **Observation 2:**   1. **SGCS Doesn't Capture Information or Capacity Loss Directly**   **SGCS only measures angular similarity between CSI matrices, ignoring magnitude differences that impact channel quality. Information loss due to quantization, feedback delay, or imperfect CSI compression is not explicitly reflected in SGCS. Capacity loss is a function of channel power and spatial correlation, but SGCS does not account for power scaling effects, which are crucial in MIMO systems.**   1. **SGCS is Weakly Correlated with Throughput**   **Throughput depends on spectral efficiency, which is determined by the achievable rate formula. SGCS does not directly relate to the determinant or eigenvalues of H, making it an incomplete indicator of spectral efficiency. The impact of CSI errors on MCS selection and beamforming is non-trivial, making it hard to infer actual throughput loss from SGCS degradation.**   1. **SGCS Does Not Capture CSI Feedback Imperfections Well**   **Compression artifacts: If a UE sends quantized CSI feedback, the distortion in signal space may not align with angular deviations captured by SGCS. Interference and noise effects: SGCS does not distinguish between CSI errors due to channel noise versus quantization/compression, leading to ambiguities in interpreting feedback accuracy.**  **Proposal 1: 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**   **Proposal 2: Investigate CSI similarity of input data distributions using Angular-Delay Channel Entropy ADCE as a statistical measure. This approach aims to reduce the reliance on more expensive monitoring procedures, such as those based on KPIs and ground truth availability, by initiating these procedures only when significant changes in data distribution are detected. The ADCE can be defined as , where , and is a virtual channel representation in angular/delay domain which separates multipath components into resolvable delay taps and spatial angles. ADCE captures the sparsity or randomness of these components.**  **Proposal 3: RAN4 to investigate defining metrics for Monitoring CSI Input Distribution to Trigger Monitoring Procedures and their Application to Dynamic CSI Compression**  **Proposal 4: For Monitoring CSI AI/ML Models investigate embedding a subset of known values, referred to as redundancy elements, into the CSI vector at the UE before encoding. These elements act as reference points that can be tracked throughout the encoding and decoding processes. The approach is designed to facilitate the monitoring of AI/ML models by providing a measurable and consistent benchmark for evaluating model performance.**  **Proposal 5: We propose a two-stage approach for CSI model monitoring, first distinguishing whether performance degradation stems from the AI/ML model or external factors and then isolating the specific issue within the model if identified as the cause.**  **Proposal 6: RAN4 to investigate a performance monitoring framework to detect issues such as data drift, model degradation, and AI/ML model aging, triggering actions like re-training, re-configuration, or model switching. The system will ensure high detection accuracy, real-time monitoring, and proactive predictions, enabling reliable adaptation to changing channel conditions while maintaining optimal performance.**  **Proposal 7: For a performance monitoring framework, event detection for data distribution drift and anomalous behavior is based on the output of a trained model that predicts performance deviations (the event). Consider the following cases as shown below:**  **(1) Indirect Event Detection: AI/ML model for SGCS EST is trained to predict SGCS from latent space c. Subsequently a detection logic can trigger event detection (2) Direct Event Detection is based on AI/ML model to directly predict an event**  **Proposal 8: Consider a new KPI for CSI compression that could offer an indication of spectral efficiency loss due to misalignment of precoders. It can be conceptually described as where Vc is the corrupted reconstructed CSI**    **Proposal 9: Investigate the benefits of defining a new KPI for CSI compression that is linked to capacity loss. For an SVD receiver this metric is approximated as , where is defined as the inter-stream interference to layer k caused by the mismatch between and V. Also is the SNR and is the eigenvalue of the channel**  **For rank 1: the proposed metric becomes , while . Under ideal conditions the term due to the orthogonality. But under imperfection this term contributes to the denominator and the capacity decreases.**  **Proposal 10: Investigate the feasibility of using the spectral efficiency KPI to determine the CQI adjustment for option 1b.** |
| R4-2503772 | CMCC | ***Proposal 1: RAN4 requirements are specified at least based on functionality (funtionality refer to AI/ML-enabled Feature/FG)***  ***Proposal 2: it is proposed that different configurations/parameters are considered for requirements/tests definition.*** |
| R4-2504025 | Vivo | **Proposal 1: For Pre-activation functionality/model update testing, at least the following issues need to be taken into account**  **Observation 1: For one sided model, taking BM as the example, regarding the impact on the possible disalignment between the metric used for performance testing and the metric used for monitoring, it would be minimal by considering the following reasons:**   * **Based on existing simulation results, beam prediction accuracy and L1-RSRP difference accuracy are highly correlated, meaning that beam prediction accuracy alone is already sufficient to reflect the model’s predictive performance.** * **Considering there is still the case that the prediction results support reporting only the beam ID (without the corresponding RSRP), Under this case, the metric used for performance testing and the metric used for monitoring can be aligned**   **Observation 2: For two-sided model, the causes for performance degradation (e.g., UE sided cause, NW sided cause, or data drift, etc.) can be identified by conducting performance monitoring**  **Obsveration 3: Compared with Option 1, with consideration of the following factors, Option 2 ensures a balanced trade-off between validation costing and practical deployment efficiency**   * **Lower Complexity – Performance monitoring leverages existing mechanisms, reducing additional implementation overhead.** * **Higher Feasibility – No test environment/test duration restrictions, e.g. laboratory testing /GCF certification.** * **Cost Efficiency – Avoids extra testing while ensuring reliable post-deployment model validation.**   **Proposal 2: By leveraging the follow factors listed in Obsveration 3, RAN4 to consider verifying AI performance after device deployment by designing the test for performance monitoring in RAN4** |
| R4-2504215 | Nokia | **Proposal 1: RAN4 to wait RAN2/RAN1 progress on LCM definitions and procedures for the two-sided CSI compression use case.**  Proposal 2: For 2-sided model case, RAN4 to define RRM requirements for LCM, prioritizing Activation, Deactivation, Switching, Fallback to non-AI operation, and Performance monitoring. |
| R4-2504320 | Huawei, HiSilicon | ***Proposal 1:*** Regarding to relation to legacy requirements in AI-CSI compression, RAN4 to discuss based on the following two options.   * Option 1: With CSI feedback reduction rate of X%, AI-based throughput gain is no worse than that of Rel-16 eType II. * Option 2: With no larger CSI feedback payload, AI-based throughput is Y% higher than that of Rel-16 eType II.   ***Proposal 2:*** Regarding to relation to legacy requirements in AI-BM, RAN4 to reuse legacy requirement for L1-RSRP for NW-side model. For UE-side model, RAN4 to discuss how to define the requirement of prediction accuracy.  ***Proposal 3:*** Regarding to relation to legacy requirements in AI-Pos, for case 2a and 3a, RAN4 to reuse legacy requirement for timing information and/or LoS/NLoS indicator if AI-based reporting and non-AI-based reporting cannot be distinguished from specification perspective.  ***Proposal 4:*** Regarding to relation to legacy requirements in AI-Pos, RAN4 to reuse legacy requirement for timing information, paired timing information and power information in case 2b and 3b.  ***Proposal 5:*** For post-deployment validation due to model change/drift, RAN4 to discuss the verification method based on the following two options.   * Option 1: Rely on model monitoring in LCM procedure. * Option 2: Define reference model to switch if newly updated model is worse than the reference model. |
|  |  |  |

## Open issues summary

### 1.2.1 Sub-topic 1-1 LCM requirements

This sub-topic captures discussions on LCM related requirements. Based on the following proposal from Nokia, it is proposed to focus on LCM for the one-sided models that are part of the WI for this meeting.

* Wait for RAN1 progress before discussing 2-sided LCM definitions and procedures (Nokia proposal 1)

Agreement from RAN4#113

**Agreement:**

* The discussion should consider use-cases specific aspects

**Agreement:**

* Whether transition from AI to non-AI would be UE initiated or network initiated
  + This will be a decision for other groups
  + RAN4 to assess what RRM requirements are needed for each case.

**Issue 1-1-1: Whether to create delay requirements for reporting/monitoring**

* Proposals
  + Option 1: Yes (CAICT proposal 2, Nokia for UE sided monitoring, proposal 19)
  + Option 2: Yes, requirements for model monitoring input data (Apple proposal 20)
  + Option 3: For NW sided monitoring, follow legacy requirements for monitoring RS resources (Nokia proposal 18)
  + Yes for 2-sided (but discuss after more RAN1/2 progress) (Nokia proposals 1-2)
* Recommended WF
  + Check if following is agreeable:
    - Create delay requirements for UE sided reporting/monitoring
    - For NW sided monitoring, follow legacy requirements for monitoring RS resources

**Issue 1-1-2: Whether to create accuracy requirements for reporting/monitoring**

* Proposals
  + Option 1: Yes (Apple proposal 19)
  + Option 2: Do not mix up tolerance and accuracy. Tolerance is the level of performance degradation before an LCM action is taken. (Nokia proposal 20-21)
  + Option 3: Yes for 2-sided (but discuss after more RAN1/2 progress) (Nokia proposals 1-2)
* Recommended WF
  + Check if it agreeable to create accuracy requirements for UE sided reporting/monitoring

**Issue 1-1-3: Whether to create delay requirements for functionality activation/switching/fallback**

* Proposals
  + Option 1: Yes, considering: (CAICT proposal 3)
    - NW decision, NW initiated
    - NW decision, UE initiaed
    - UE decision/reporting, NW initiated
  + Option 2: UE initiated transition may not be feasible if UE side proprietary information is involved (Huawei proposal 1)
  + Option 3: No activation / switching delay requirements for CSI in RAN4 (Qualcomm proposal 1)
  + Option 4: Activation delay requirements at least for periodic CSI reporting, FFS for aperiodic and semi-persistent (Ericsson proposal 8)
  + Option 5: For positioning, firstly wait for progress on case 1 before deciding on activation delay (Qualcomm proposal 2)
  + Option 6: No need for a requirement for positioning case 1 (Ericsson proposal 7)
  + Option 7: For periodic and sem-persistant setB resources, wait for more clarity from RAN1 (Qualcomm proposal 3)
  + Option 8: For aperiodic setB resources, define delay requirements (Qualcomm proposal 4)
    - The timeline starts when UE receives the network command that triggers the of setB resources. The timeline ends when UE is expected to convey the inferred properties of setA resources via uplink.
  + Option 9: For CSI periodic, delay requirement is from RRCReconfiguraitonComplete message and for semi-persistant/aperiodic CSI, from receiving MAC-CE or DCI indication (CMCC proposal 1, Nokia proposals 24-25)
  + Option 10: The ending point of the AI/ML functionality activation delay is when UE starts reporting inference to the network. (Nokia proposal 26)
  + Option 11: Do not define deactivation requirements for beam management (Qualcomm proposal 5)
  + Option 12: Define deactivation requirements for beam management (Ericsosn proposal 9, Huawei for NW initiated cases Proposal 2)
  + Option 13: The starting point for deactivation delay is when the UE receives the command to commence deactivation. The endpoint is FFS, depends on RAN1/2 agreements (Nokia proposal 27-28)
  + Option 14: FFS whether fallback involves more than deactivation delay (Ericsson proposal 10 for BM, Huawei proposal 3)
  + Option 15: Create the following definition for functionality switching (Ericsson proposal 10):
    - **AI/ML functionality switching** is a situation when one AI/ML functionality is deactivated and another functionality is activated by the same message.
  + Option 16: Create the following definition for functionality fallback (Ericsson proposal 11):
    - **Fallback** from AI/ML operation to non-AI/ML operation is a situation when one or more AI/ML functionalities are deactivated and there is no relevant activated AI/ML functionality left.
  + Option 17: Create the following definition for functionality activation (Nokia proposal 22):
    - AI/ML functionality activation refers to the process of enabling an applicable functionality to perform inference
  + Option 18: Create the following definition for functionality de-activation (Nokia proposal 23):
    - functionality deactivation refers to the process of network deactivating an active functionality
  + Option 19: Delay requirements should be use case specific (Nokia proposal 29):
  + Yes for 2-sided (but discuss after more RAN1/2 progress) (Nokia proposals 1-2)
* Recommended WF
  + Check if the following definitions can be agreed:
    - **AI/ML functionality switching** is a situation when one AI/ML functionality is deactivated and another functionality is activated by the same message.
    - **Fallback** from AI/ML operation to non-AI/ML operation is a situation when one or more AI/ML functionalities are deactivated and there is no relevant activated AI/ML functionality left.
    - **AI/ML functionality activation** refers to the process of enabling an applicable functionality to perform inference
    - **functionality deactivation** refers to the process of network deactivating an active functionality
  + For deactivation:
    - The starting point is when the UE receives the command to commence deactivation
    - FFS when the endpoint for deactivation is
    - FFS whether fallback delay is the same as deactivation delay or has some other components
  + For CSI, discuss:
    - Option 1: For CSI periodic, delay requirement is from RRCReconfiguraitonComplete message and for semi-persistent/aperiodic CSI, from receiving MAC-CE or DCI indication
    - Option 2: No activation/switching related requirements for CSI
  + For beam management, discuss:
    - For periodic and semi-persistant sebB resources, FFS
    - For aperiodic setB resources, define activation delay requirement
      * The timeline starts when UE receives the network command that triggers the of setB resources. The timeline ends when UE is expected to convey the inferred properties of setA resources via uplink.
    - For deactivation:
      * Option 1: No requirement
      * Option 2: Define a requirement
  + For positioning:
    - For case 1:
      * Option 1: Reach decision no delay requirement
      * Option 2: Wait until further progress on whether to define a performance requirement
  + Regarding which circumstances to apply delay requirements, discuss:
    - NW decision, NW initiated
    - NW decision, UE initiaed
    - UE decision/reporting, NW initiated

**Issue 1-1-4: Whether to create delay requirements for data collection**

* Proposals
  + Option 1: Yes (Ericsson proposal 11)
  + Option 2: TBA
* Recommended WF
  + TBA

**Issue 1-1-5: Whether to create delay requirements for model switching/fallback**

Note that this proposal is for model switching, not functionality switching.

* Proposals
  + Option 1: Yes (CAICT proposal 2)
* Recommended WF
  + TBA

**Issue 1-1-6: Monitoring report metric**

* Proposals
  + Option 1: For CSI prediction, intermediate KPI (CMCC proposal 2)
  + Option 2: Metrics based on CSI input distribution (Apple proposal 3)
  + Option 3: For 2-sided CSI, embed a subset of know values to enable tracking (Apple proposal 4)
  + For 2-sided CSI, devise a means to devise whether the degradation is from an external factor or the model, and then the specific model issue (Apple proposal 5)
  + Use a trained model to predict performance deviations due to data drift (Apple proposal 7)
  + For 2-sided CSI, consider a metric relating to the spectral efficiency (Apple proposals 8-10)
  + Wait for RAN1 progress before discussing 2-sided LCM definitions and procedures (Nokia proposal 1)
* Recommended WF
  + RAN1 is discussing the monitoring metric. Check if RAN4 should send any feedback based on testing feasibility.

### 1.2.2 Sub-topic 1-2 Generalization testing

**Issue 1-2-1: What needs to be included for generalization testing**

* Proposals
  + Option 1: At least different propagation conditions and scenarios, UE parameters, gNB array parameters. Downselect to make a benchmark valid across a range of unseen scenarios (Apple proposal 2)
  + Option 2: For downselection, a study of AI model’s behavior changes with different scenarios and conditions is needed (Apple proposal 5, Nokia proposal 33)
  + Discuss the number of requirements in the requirement set during the performance phase (CATT proposal 4)
  + For CSI, consider at least antenna port layouts, antenna spacing, antenna virtualization, carrier frequencies, outdoor/indoor distributions, UE speeds, bandwidths, feedback granularities, CSI feedback payloads (Apple proposal 1)
* Recommended WF
  + Discus whether the following can be agreed:
    - For generalization testing, consider the following list of potential variations:
      * Deployment conditions (Carrier frequency, cell type, indoor/outdoor etc.)
      * Propagation conditions
      * gNB array parameters (port layouts, antenna spacing, antenna virtualization etc.)
      * UE parameters (e.g. feedback granularity)
      * Further aspects to be added
    - Reduce to a manageable set of tests to benchmark across unseen scenario
    - Study the AI model behaviour changes across different scenarios and conditions

**Issue 1-2-2: Relation between generalization testing and feature configurations/capability signalling**

* Proposals
  + Option 1: Identify generalization tests based on capability signalling (Apple proposal 9, 12)
  + Define scenarios per use case in a future release (Apple proposal 11)
  + “other scenarios and conditions” are scenarios or configurations not reported by UE capability signaling (Apple proposal 13)
  + Decide after RAN1 achieve agreements on UE capability (CATT proposal 3)
  + At least one test per specific sub-set of applicable feature configurations (Nokia proposal 30, CMCC proposal 1,2)
  + There should be a minimum set of configurations (Nokia proposal 31)
  + If there are legacy configurations/parameters, at least these should be included (Huawei proposal 4)
* Recommended WF
  + For discussion:
  + Define at least one requirement/test per applicable feature configuration.
    - Call the requirement/test per applicable feature configuration a “requirement set”. A UE must pass all requirements in a “requirement set”, but which requirement sets are applicable is FFS.
    - Options for the relation between the “requirement set” and capability signalling/associated IDs
      * Option 1: One requirement set per UE capability
      * Option 2: One requirement set per network additional condition
      * Option 3: Option1+option2
      * Option 4: Legacy configurations/parameters (where available) shall be in at least one mandatory requirement set.
      * Option 5: other options not precluded
    - FFS whether each “requirement set” consists of a single requirement or several requirements
      * Option 1: One requirement per “requirement set”
      * Option 2: Several requirements per “requirement set” to cover the range of conditions expected within the scenarios covered by the requirement set.
      * Option 3: Other options not precluded.
      * May be use-case and requirement set dependent

**Issue 1-2-3: Relation between generalization testing and additional conditions signalling**

* Proposals
  + Option 1: Ensure consistency between additional conditions signalling and testing (Apple proposal 10, 17)
  + No performance requirements that imply a change in network configuration or associated ID during the test (Nokia proposal 32)
* Recommended WF
  + Check if the following is agreeable:
    - There shall be consistency between applicable conditions signalling and testing
    - No test will be defined that implies a change of network condition or associated ID during the test.

**Issue 1-2-4: Non-static scenarios**

* Proposals
  + Option 1: Include semi-static in generalization tests, but only if static are not enough (Apple proposal 3-4, Nokia proposal 34)
  + Consider to create a mixed dataset containing samples from different configurations in order to test generalization (Apple proposal 7)
* Recommended WF
  + Check if the following can be agreed:
    - Static scenarios are assumed by default. Non-static scenarios are introduced on a use-case specific basis if needed for testing.
    - FFS whether a mixed datset can be created for testing generalization, and whether such a mixed dataset would be a static or non-static scenario.

### 1.2.2 Sub-topic 1-3 Post-deployment testing

In this sub-topic, post-deployment monitoring is further discussed

Agreement from RAN4#112bis:

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

When operating in the field, two aspects of UE operation may impact performance:

* Update or fine tuning or addition/removal of models, if applicable, resulting in a change of functionality which may impact functionality performance
* Data drift / mismatch between the conditions encountered by the UE in the field and the training data, which may impact functionality performance.

For dealing with potential changes in the performance of functionalities two options may be available. The options are not mutually exclusive:

* **Option 1**: Conduct the validation of a change in ~~AI model/~~functionality before its deployment/activation in already deployed UEs
  + Validation takes into account the UE hardware in which the model is to be deployed/activated.
  + FFS whether the validation takes place in the field (i.e., on each individual device) or in the lab conditions (i.e., per device type/model)
  + FFS whether validation can be done at the device along with inference for another active functionality(ies)
  + FFS on the feasibility
  + FFS on one possibility for consideration for option 1 is to capture model input (and if needed other test data such as ground truth) during conformance testing. This stored data can later be used to validate new or updated models.
  + Other aspects not precluded.
* **Option 2**: Using performance monitoring and LCM procedures
  + Performance monitoring will be designed in other groups
  + RAN4 may consider the need and feasibility of requirements and tests to ensure consistency and accuracy of monitoring metrics or other monitoring related data sent from the UE, and set requirements as feasible/needed.
  + FFS on monitoring can be used for managing fallback, changes in functionality, model update/model switching/model transfer, if applicable

FFS: RAN4 needs to clarify whether changed functionalities that did not pass conformance testing are expected to be activated at the device.

For dealing with drift / mismatch between the conditions encountered by the UE in the field and the training data for the model, which may impact functionality performance, monitoring is needed.

* + Performance monitoring will be designed in other groups
  + RAN4 may consider the need and feasibility of requirements and tests to ensure consistency and accuracy of monitoring metrics or other monitoring related data sent from the UE, and set requirements as feasible/needed.
  + Monitoring can be used for managing fallback, change in functionality, model update/model switching/model transfer, if applicable

**WF from RAN4#114:**

|  |  |  |
| --- | --- | --- |
|  | Option 1 | Option 2 |
| New name | Pre-activation functionality/model update testing | Post deployment management based on LCM |
| Description | Conduct the validation of a change in AI functionality before its deployment/activation in already deployed UEs  • Validation takes into account the UE hardware in which the model is to be deployed/activated. | Using performance monitoring and LCM procedures  • Performance monitoring will be designed in other groups  • RAN4 may consider the need and feasibility of requirements and tests to ensure consistency and accuracy of monitoring metrics or other monitoring related data sent from the UE, and set requirements as feasible/needed. |
| Sub-options | Possible collection of input data to a model during conformance testing for use later on. |  |
| Further potential clarifications (not resolved) | The following proposals have been made for option 1 but not discussed.  o Whether the testing of a new model is on a device or in a lab.  o Whether models that have been updated post conformancetesting should be kept in a device.  o Whether parallel operating of one model and (in-device) testing of another can be assumed.  o Whether signaling from a UE of changes/updates are needed.   * Whether there is any relation to GCF procedures. * Whether it is possible to differentiate “major” and “minor” model changes and only apply option 1 to major changes.   + Definition of Fine-tuning and Model update     - Model update can have a performance impact.     - Fine-tuning might have little performance impact. * Hybrid Validation approach   + A hybrid approach integrating both pre-activation functionality/model update testing (Option 1) and post-deployment management based on LCM (Option 2) aims to provide a balanced validation strategy. | The following proposals have been made for option 1 but not discussed.   * Whether the metric used for performance testing and the metric used for monitoring can be aligned. * Whether monitoring used for post-deployment testing should be NW sided only. |
| Significant issues | Whether the option is reasonable or would cause too much complexity and test time burden when introducing updates. | Whether LCM monitoring can actually unambiguously identify individual model performance with reasonable complexity (considering other variations due to e.g. channel, interference, scheduling, other AI models etc.)  Whether LCM monitoring can have RAN4 requirements to ensure consistent monitoring between different UEs. |
|  |  |  |

**Issue 1-3-1: Naming of options and other vocabulary**

* Proposals
  + Option 1: Rename “option 1” as “Post-deployment pre-activation functionality/configuration update testing” (Nokia proposal 2)
    - “Pre- activation” means “the functionality/configuration is available in (and supported by) the UE device, and it is not RRC/MAC activated” (Nokia proposal 3)
  + Option 2: Rename “option 2” as “Post-deployment post-activation functionality testing based on performance monitoring” (Nokia proposal 4)
  + Option 3: “Test active” means “the functionality/ configuration which is available in (supported by the) UE, it is RRC configured by the gNB/NW for post-deployment testing and, has been RRC/MAC activated by the gNB/NW for post-deployment testing” (Nokia proposal 5)
* Recommended WF
  + Check if agreeable

**Issue 1-3-2: Prioritization of options**

* Proposals
  + Option 1: Prioritize option 2 (Post deployment management based on LCM) (CAICT Proposal 1, Vivo proposal 2)
  + Option 2: Determine whether monitoring is useful and interoperable before dropping the first option (Ericsson proposal 4)
  + Option 3: Combination of option 1 and option 2 (Nokia proposal 17)
  + Option 4: Performance monitoring and applicability can handle post-deployment (CATT proposal 1)
  + Option 5: More clarification needed before pursuing option 1 (CATT proposal 2)
  + Option 6: Either option 2, and/or define a reference model to switch to if the new model is worse than the reference model (Huawei proposal 5)
* Recommended WF
  + Further elaborate monitoring (option 2) to check whether it is sufficient
  + Further elaborate what is meant by option 1; is it on-UE testing or a validation statement when a model is updated based on e.g. simulating performance, or something else
  + Recognize general preference for option 2, but decide after the above two points are further progressed.

**Issue 1-3-3: Proposal for new OPTION 3:**

* Proposals
  + Option 1: The network may have an ID based on training data collection. If the UE lacks a model with the ID, the network transfers it. If the network does not have the ID, the UE can fine tune based on the NW collecting training data and a new ID is assigned, related to an additional condition or fallback can take place if there is no ID and no training data (Apple proposals 22-26).
  + Option 2: To detect drift, use a server for device side monitoring and database management(Apple proposal 27-28)
  + Option 3: The network can trigger fine tuning at the network or UE side (Apple proposals 29-32)
* Recommended Discussion
  + Questions to discuss:
    - The proposed new option 3 involves use of signaling, model IDs and training/fine tuning to deal with performance. But the basic mechanisms of detecting performance seem the same as options 1 and 2 ? And then the model management/fine tuning etc. may be in other WG scope ?
    - Compared to OPTION 1-2, what is the additional RAN4 related requirements/discussion for OPTION 3 ?

**Issue 1-3-4: Definition of fine tuning and model update**

* Proposals
  + Ericssoon proposal 3 is to discuss, but there is no specific solution
  + Option 1: Define as follows (KTL Proposal 2, 3):
    - Fine tuning is adjustment of (a sub-set of) model weights without changing the structure
    - Model update is full replacement of an AI/ML model. This may include different architecture.
  + Option 2: May be difficult to standardize. Device could decide, but unclear if it can be useful or objective (Vivo)
* Recommended WF
  + Discuss whether classification of fine tuning and model update is seen as potentially feasible or not

**Issue 1-3-5: Location for post-deployment testing option 1**

* Proposals
  + Option 1: At the place where the update is created (Ericsson proposal 1)
  + Option 2: In the UE (Nokia proposal 1, 11)
  + Option 3: Neither a good option. Lab testing too high effort, on-device testing may not capture performance caracteristics (vivo)
* Recommended WF
  + For options not in UE, elaborate what kind of testing is meant
  + For in UE testing, discuss further issue 1-2-6
  + Based on this decision, check if one of the options can be selected, or some “option 1a/option 1b” description is needed in the WF.

**Issue 1-3-6: Conditions and requirements for post-deployment testing option 1**

* Proposals
  + Option 1: Assuming that the post-deployment validation takes place in the UE with a model in “test active” mode (testing but not producing inference output for the functionality), discuss the framework and the potential for configurable conditions and requirements (Nokia proposals 6-7)
  + Option 2: Enable both active (i.e. providing inference) and test-active (i.e., under test) models to run in parallel in the UE (Nokia proposal 13)
  + Option 2: Ensure at least one model that has passed conformance testing is always available (Nokia proposal 12)
* Recommended WF
  + TBA

**Issue 1-3-7: Relation to GCF**

* Proposals
  + Option 1: Ask RAN5 to consider (Ericsson proposal 2)
  + RAN5 should discuss (not clear when) (Nokia proposal 16)
* Recommended WF
  + TBA

**Issue 1-3-8: Metric for performance monitoring**

* Proposals
  + Option 1: Use a difference metric for monitoring so that the reporting can take place when the metric changes (KTL proposal 4)
  + Option 2: Explcitly define overfitting validation in post-deployment scenarios, including procedures and signaling (KTL proposal 8)
  + Option 3: Design similar metrics for performance monitoring and conformance tests (Xiaomi proposal 3)
* Recommended WF
  + TBA

**Issue 1-3-9: Monitoring ability to identify issues**

* Proposals
  + Option 1: Cannot and should not identify issues in particular UE model (Nokia proposal 9)
  + Option 2: For two-sided, the cause can be identified (NW side, UE side or data drift) (Vivo)
* Recommended WF
  + TBA

**Issue 1-3-10: Requirenents for performance monitoring**

* Proposals
  + Option 1: Consider the need and feasibility of requirements/tests for monitoring (Nokia proposal 8, 10)
  + Option 2: Define requirements for performance monitoring (Apple proposal 19)
* Recommended WF
  + See issues 1-1-1, 1-1-2

**Issue 1-3-11: Signalling for post-deployment testing (Nokia proposal 14 to discuss)**

* Proposals
  + Option 1: Optimize signalling for LCM based post-deployment monitoring (KTL proposal 1)
  + Option 2: If there is network side validation, mandate signalling of UE model changes (KTL proposal 5)
  + Option 3: RAN4 should discuss and define the post-deployment monitoring procedure, and then RAN1/2 define the appropriate signaling. (KTL proposal 6)
  + Option 4: For “option 2” (LCM based post-deployment), signalling will be needed (Nokia proposal 15)
* Recommended WF
  + TBA

### 1.2.3 Sub-topic 1-4 Other

**Issue 1-4-1: Relation to legacy requirement**

RAN4#112bis WF:

Agreement:

* For CSI prediction, AI/ML based performance requirement should not be worse than the legacy (non-AI/ML based) if a relevant equivalent legacy requirement/feature exists.
  + FFS which legacy requirement/feature are relevant
  + FFS whether AI/ML based performance should be better than the legacy
* FFS on target for AI/ML based beam management
* Proposals
  + If the AI configurations/parameters are the same as legacy, take legacy as starting point (Huawei proposal 4)
  + If a new test setup is proposed for AI, discuss further whether to determine non-AI requirements in each case so that AI can have more stringent requirements, or just directly set absolute AI requirement (Huawei proposal 5)
  + For beam management, AI/ML should not be worse than legacy (CMCC proposal 3, Huawei proposal 2 for NW sided)
  + For RSRP accuracy, measurement accuracy my have an impact on whether AI requirements can exceed legacy (Xiaomi proposal 1)
  + Beam ID prediction accuracy is an example where there is no performance requirement in “legacy” (Xiaomi proposal 2)
  + For positioning, use legacy requirements for case 3a/b (Huawei proposals 3-4)
  + For 2-sided, either lower feedback rate or higher throughput than legacy (Huawei proposal 1)
* Recommended WF
  + Check if the following can be agreed (Note that this is only really a refinement of the existing agreement):
    - If “legacy” requirements are applicable for a functionality, AI performance ideally should be no worse and may be better.
      * For beam management, if there is an RSRP accuracy requirement, measurement error may mean that AI does not exceed “legacy”. This should be discussed further in the beam management thread.
      * FFS whether “better than legacy” means improved throughput for the same overhead or lower overhead for the same throughput; this should be discussed in the use-case threads.
    - If there is no legacy requirement:
      * Option 1: Define AI requirement without considering legacy
      * Option 2: Study non-AI performance in the same conditions as the AI requirement.
      * (Check in meeting if one option could be decided or not)

**Issue 1-4-2: Side conditions**

* Proposals
  + Side conditions should be the same for legacy and AI/ML (Apple proposal 18)
* Recommended WF
  + TBA

**Issue 1-4-3: CSI input data distribution**

* Proposals
  + For CSI, investigate use of ADCE as a means for investigating CSI distribution (Apple proposal 2)
* Recommended WF

TBA

**Issue 1-4-4: Study of usefulness of synthetic test data when testing models trained on real data**

Agreement from RAN4#112

Agreement:

* Use synthetic channels as the default assumption for conformance requirements/test.
  + RAN4 may 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.
    - The field data can be considered for the analysis.
* Proposals
  + Option 1: Do an analysis per use case to determine the reliability of using synthetic channels (Apple proposal 33)
* Recommended WF
  + The proposal is already captured in the previous agreement.

**Issue 1-4-5: Testing goal**

Agreement from RAN4#112:

Agreement:

* The testing goal is to verify whether the minimum performance of AI/ML functionality/feature can be achieved
  + FFS on whether and how many different test configurations/parameters are used for tests.
  + LCM would be tested
* Proposals
  + Option 1: Focus on “option 2” (Apple proposal 1)
* Recommended WF
  + “Option 2” is from a previous agreement and corresponds to the agreement at RAN4#112.

# Topic #2: CSI compression two sided models

This topic addresses the study on two-sided CSI compression. It is divided into two sub-topics. The first sub-topic is about the simulation campaign investigating options 3 and 4 for requirement definition. The second sub-topic addresses issues relating to options 3 and 4 that are not directly related to the simulations.

## 2.1 Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| R4-2503255 | Qualcomm | **Observation 1: The SGCS obtained with separate encoder decoder training (QC encoder with Samsung’s frozen decoder) in option 3 with track 2 mixed dataset is almost identical to the SGCS obtained with Samsung’s joint encoder-decoder training.**  **Observation 2: Option 3 (fully specified decoder) leads to good SGCS in, at least, the scenarios where encoder is trained with the same dataset as the decoder.**  **Observation 3: Option 4a leads to good SGCS when Qualcomm’s “own” encoder, derived through a two-step approach from the labelled training dataset, is tested with Samsung decoder and mixed dataset. The feasibility of option 4a will depend on how Qualcomm’s “own” encoder performs other companies’ “nominal” decoders that were generated based on labelled dataset.**   * + - * **Note 1: “Labelled” dataset refers to (encoder input, encoder output)**       * **Note 2: “Labelled dataset” was generated based on mixed dataset and Samsung reference encoder.**   **Observation 4: In option 4b,**   * **TE is expected to train its decoder based on reference encoder.** * **UE is expected to develop a nominal decoder that works well with the reference encoder and then train its encoder based on the nominal decoder.** * **There will be mismatch between TE and UE if the training dataset, i.e., channel information, is not specified.**   **Observation 5: Training dataset, i.e., channel information, and reference encoder both need to be specified in option 4b.**  **Observation 6: If channel information and reference encoder are specified in option 4b, then option 4b becomes a superset of option 4a where (channel information, encoder output) is specified.**  **Proposal 1: RAN4 specifies both training dataset (channel information) and reference encoder in option 4 and focuses *only on this option* for the partially specified approach.**  **Proposal 2: RAN4 also considers specifying a reference decoder in option 4 so that UE can derive its encoder through one step and TE can get a guideline to implement its decoder during the tests.** |
| R4-2503258 | NTU | **Proposal 1: Clarify the following for procedures of option 4a test decoder feasibility study:**   * **What is the content of dataset from option 3 track 1 and option 3 track 2?** * **Whether option 3 track 1 has an encoder selected besides the selected fixed decoder** * **If we use the encoder/decoder selected from option 3 track 1, what’s the additional alignment needed given that we agreed to have “*Each company is encouraged to train own encoder using the fixed decoder and their own dataset, and then test with every companies datasets, in order to check for consistency in results*” procedure in option 3 discussion, which is the same as step 6 and 7. Similar issue is observed if we use dataset from option 3 track 2** * **If the mixed dataset from track 2 is used, how to train the encoder and decoder based on it given that it only contains encoder input?** * **How to verify whether the following agreed aspects of option 4 are feasible or not?**   + ***Test repeatability should be ensured (variation among TE vendor implementations should be bound)***   + ***Other vendors should also be able to develop such a decoder and which can deliver similar performance***   **Proposal 2: Clarify the following for procedures of option 4b test decoder feasibility study**   * **How to select one encoder from companies’ own encoder in option 3 track 1?** * **Given that option 4b aims to select one reference encoder, why we need to check multiple encoder alignment in step 6 after one is selected already?** * **How to verify whether the following agreed aspects of option 4 are feasible or not?**   + ***Test repeatability should be ensured (variation among TE vendor implementations should be bound)***   + ***Other vendors should also be able to develop such a decoder and which can deliver similar performance***   More importantly, we want to ask for clarification what kind of feasibility is verified by the procedures captured in the WF.  **Proposal 3: Clarify the feasibility that can be verified by the agreed procedures and its limitation as follows:**   * **Option 3 track 1: We verify the performance alignment of the selected test decoder and its own dataset is feasible. However, if we select a different decoder with a different dataset in the WI stage, the performance alignment may not be feasible, and RAN4 has to redo the feasibility study.** * **Option 3 track 2: We verify the performance alignment of the selected test decoder and the selected multiple source datasets is feasible. However, if we select a different decoder with different datasets in the WI stage, the performance alignment may not be feasible, and RAN4 has to redo the feasibility study.** * **Option 4a: We verify the performance alignment of test decoders trained with the selected dataset. However, if we select a different dataset in the WI stage, the performance alignment may not be feasible, and RAN4 has to redo the feasibility study.** * **Option 4b: We verify the performance alignment of test decoders trained with the selected reference encoder. However, if we select a different reference encoder in the WI stage, the performance alignment may not be feasible, and RAN4 has to redo the feasibility study.**   **Proposal 4: Additional steps for option 4a and 4b feasibility study:**  **Define the decoder verification criterion from one of the following options (for 4a and 4b, respectively):**  **(Option 4a: specify dataset)**   * **4a-1: RAN4 derive the threshold via simulation study (setup following option 3 discussion), and check whether it’s feasible to derive a test decoder satisfying**     **where is the set of decoder inputs (latent messages) in the specified dataset (from step 2), and *f* is the chosen loss/similarity function is the test decoder output and is from the dataset.**   * **4a-2: RAN4 studies whether it is feasible to derive a test decoder satisfying**   **where the notations are the same as 4a-1, and is the test decoder output space.**  **(Option 4b: specify reference encoder)**   * **4b-1: RAN4 derive the threshold and the encoder input *X* generation procedure via simulation study (setup following option 3 discussion), and check whether it’s feasible to derive a test decoder satisfying**     **where is the reference encoder output (latent message) of encoder input *x*, and is the sampled encoder input in the test decoder verification procedures. Note that RAN4 needs to specified the generation procedures of *X* and specifying**   * **4b-2: RAN4 studies whether it is feasible to derive a test decoder satisfying**   **where is the test decoder output space and *Z* is the reference encoder output space.**  **RAN4 check if the decoders derived from the above procedure can satisfy test repeatability requirement, i.e., the loss function delta between two different decoders when connecting to the same encoder, is within a RAN4 agreed margin. If yes, we can conclude that option 4 (at least one sub-option) is feasible.**  **Proposal 5: Next steps for feasibility study of option 4c**   * + - 1. **Use the encoder input part of the dataset from option 3 feasibility study, use principle component analysis to derive eigenvectors and eigenvalues of the encoder input dataset.**       2. **RAN4 decides how many dominant eigenvectors, and the number of quantization bits for the coefficients of eigenvectors (sum to the number of bits in the latent message) to include in indexing procedure in proposal 3.**       3. **Companies propose any encoder-decoder pair, with the latent message reordering based on the procedures in proposal 3 and the dominant eigenvectors decided in step 1**       4. **RAN4 check if the decoders derived from the above procedure can satisfy test repeatability requirement, i.e., the loss function delta between two different decoders when connecting to the same encoder (also following the latent message reordering procedure), is within a RAN4 agreed margin. If yes, we can conclude that option 4c is feasible.**   **Proposal 6: The latent message reordering procedure is in the following:**  **Step 1: Perform principle component analysis (PCA) on encoder input dataset, i.e., derive the SVD of the covariance matrix of the encoder input dataset.**  **Step 2: Select the eigenvectors corresponding to the *M* largest eigenvalues as dominant eigenvectors.**  **Step 3: Derive the coefficients of each of the decoder output on the (incomplete) basis composed of the selected *M* eigenvectors.**  **Step 4: Suppose the latent message length is *N*. Distribute the *N* bits to represent the coefficients of the *M* eigenvalues proportional to the magnitude of the corresponding eigenvalues.**  **Step 5: Quantize the coefficients to form the latent message corresponding to each decoder output. The mapping from the new latent message from these steps to the original latent message before reordering is added before decoder input. The mapping from the original encoder output to the new latent message from these latent messages is added to the encoder output.**  **Proposal 7: Additional margin to account for the mismatch need to be considered when determining requirement.**  **Proposal 8: Consider the test decoder selection criterion based on the agreed step, i.e., check whether performance alignment across contributing companies’ “own encoder”, which may have different structures, can be achieved.  Therefore, companies can bring the test decoder proposal together with performance metrics achieved when connecting to different encoders with different structures, e.g., the structure options for test decoder discussion, but trained with the decoder input and output dataset from the proposed decoder.** |
| R4-2503271 | Mediatek | **Observation #1: Option 3 Track 1 SGCS results with frozen pretrained decoders using companies own datasets can in most cases closely match reference SGCS results where Encoder and Decoder are trained together (Reference).**  **Observation #2: Dataset can have an impact on SGCS results for some shared frozen pretrained decoder models.**  **Observation #3: Option 3 Track 2 SGCS results with frozen pretrained decoder (trained with mixed dataset) using own or mixed dataset can in most cases closely match reference SGCS results where Encoder and Decoder are trained together (Reference).**  **Observation #4: Option 3 Track 2 SGCS results with frozen pretrained decoder (trained with mixed dataset) using own or mixed dataset can be poor in unquantized training where we suspect convergence issue with agreed hyperparameters.**  **Observation #5: Option 4a SGCS results with using labelled dataset can closely match reference SGCS results where Encoder and Decoder are trained together (Reference).**  **Observation #6: Option 4a SGCS results with frozen pretrained decoder from previous step can in most cases closely match reference SGCS results where Encoder and Decoder are trained together (Reference).**  **Observation #7: Option 4a SGCS results with frozen pretrained decoder from previous step can be poor in unquantized training where we suspect convergence issue with agreed hyperparameters.** |
| R4-2503410 | CATT | **Observation 1: In Option 3 track 1, SGCS spans from 0.6868 to 0.7246 for un-quantized model and from 0.6557 to 0.7013 for quantized model.**  **Observation 2: Option 3 track 1 is feasible when characteristic of training datasets and test datasets are aligned.**  **Observation 3: In Option 3 track 2, SGCS is 0.7206 for un-quantized model and 0.713 for quantized model.**  **Observation 4: CNN encoders with different number of ResNet blocks have slightly different SGCS.**  **Observation 5: Option 4a and 4b will be essentially the same, i.e., equivalent to Option 4a, if mixed dataset is used to train own decoder in Option 4b and own encoder in both options.**  **Observation 6: In Option 4a, SGCS is 0.7411 for un-quantized model and 0.6939 for quantized model when the mixed dataset is used; 0.7265 and 0.6722 respectively when own dataset is used.**  **Observation 7: Compared with SGCS with the mixed dataset used to own encoder, a performance degradation, around 2%, is observed when own dataset is used.**  **Observation 8: In Option 4b, SGCS is 0.749 for un-quantized model and 0.6841 for quantized model when own dataset is used.**  **Proposal 1: Use own dataset to train own decoder and own encoder in Option 4b.**  **Proposal 2: In Option 3, the encoder corresponding to the test decoder is selected as reference encoder.**  **Proposal 3: FFS where to capture the reference encoder when defining performance requirements in WI, if exists.**  **Observation 9: For both Option 3 track 1 and 2, it would be helpful to capture the training dataset.**  **Proposal 4: Storing the training dataset in a dedicated folder can be considered for both Option 3 track 1 and 2.** |
| R4-2503516 | Apple | **Observation 1: The feasibility of Option 3 Test Decoder under Track 1 cannot be established for all the chosen test decoders due to the performance dependency on the testing dataset. That is, in the absence of an agreed training dataset, if each company trains each own encoder with own training dataset, then the testing performance at TE with the trained DUT will be testing-dataset dependent. DUT will perform differently depending on the testing dataset. From simulation results, there is a statistical difference between companies’ datasets.**  **Observation 2: Performance of Own trained (mixed dataset) Encoder with Mixed Dataset trained Decoder is more consistent across various companies testing datasets. However, the performance on testing datasets that don’t belong to the training dataset (QC, CATT) is significantly degraded mirroring the issues observed in track1.**  **Observation 3: To achieve reproducible tests and guarantee interoperability, the specification will need to capture the following aspects:**   * **The latent space representation to be captured by the decoder** * **An aggregate dataset from UE vendors, featuring UE-preferred encoders in the form of {V, c}, aiming to enrich the latent space and enhance interoperability.**   **Observation 4: Input CSI data distribution can be used to detect a monitoring event. By detecting monitoring events without ground truth CSI excessive overhead signaling can be avoided**  **Observation 5:**   1. **SGCS Doesn't Capture Information or Capacity Loss Directly**   **SGCS only measures angular similarity between CSI matrices, ignoring magnitude differences that impact channel quality. Information loss due to quantization, feedback delay, or imperfect CSI compression is not explicitly reflected in SGCS. Capacity loss is a function of channel power and spatial correlation, but SGCS does not account for power scaling effects, which are crucial in MIMO systems.**   1. **SGCS is Weakly Correlated with Throughput**   **Throughput depends on spectral efficiency, which is determined by the achievable rate formula. SGCS does not directly relate to the determinant or eigenvalues of H, making it an incomplete indicator of spectral efficiency. The impact of CSI errors on MCS selection and beamforming is non-trivial, making it hard to infer actual throughput loss from SGCS degradation.**   1. **SGCS Does Not Capture CSI Feedback Imperfections Well**   **Compression artifacts: If a UE sends quantized CSI feedback, the distortion in signal space may not align with angular deviations captured by SGCS. Interference and noise effects: SGCS does not distinguish between CSI errors due to channel noise versus quantization/compression, leading to ambiguities in interpreting feedback accuracy.**  **Proposal 1: To establish the feasibility of the Test Decoder under Option 3 and Track1, each company should train its own encoder using the fully specified Test Decoder and its own training data. The trained encoder should then be tested with the Test Decoder using datasets from different companies and demonstrate performance consistency**  **Proposal 2: Investigate a Channel State Information (CSI) metric that provides a statistical measure of how well CSI can be compressed, directly translating to SGCS performance. Using this approach, a dissimilarity metric between CSI training and testing datasets can be derived to predict or interpret SGCS results. One such metric could be Angular-Delay Channel Entropy (ADCE), which quantifies the uncertainty in CSI’s spectral distribution and reflects its compressibility. The ADCE can be defined as , where , and is a virtual channel representation in angular/delay domain which separates multipath components into resolvable delay taps and spatial angles. ADCE captures the sparsity or randomness of these components.**  **Proposal 3: To determine the initial feasibility of test decoder under Option 3 and Track 2, each company to train Own Encoder with Test decoder with the mixed dataset. Subsequently each company tests the Enc/Dec pair with the mixed dataset to ensure performance consistency across all companies.**  **Proposal 4: RAN4 to investigate potential performance loss due to the dataset distribution mismatch (field data distributions) when one or both parts of the AI/ML models are specified with a synthetic dataset.**  **Proposal 5: RAN4 to investigate the similarity between Option 3 and Inter-vendor collaboration direction C (Fully standardized reference model(s) and parameters with specified CSI generation part and/or CSI reconstruction part (Inter vendor collaboration option 1) from RAN1 study**  **Proposal 6: RAN4 to investigate how many test decoders need to be specified under Option 3. Identify the granularity of scenarios/deployments that will dictate a new test decoder.**  **Proposal 7: To complete feasibility analysis under Option 3, the following question should be answered:**  **Performance Consistency Across UE Encoder Implementations: Can the performance be maintained across different UE encoder implementations? Specifically, if the UE encoder uses a different AI/ML backbone (e.g., a Transformer-based architecture), does the system still achieve comparable results?**  **Proposal 8: To align with real-world deployment scenarios and advance toward defining performance requirements, the feasibility analysis should be expanded to include at least two layers**  **Proposal 9: RAN4 to employ the procedures shown in the flowchart for studying the feasibility analysis of test decoder Option 3. (It captures both Track-1 and Track-2)**  **Proposal 10: The following criteria should establish the feasibility of the test decoder:**   1. **The test decoder (along with the encoder) should deliver reasonable throughput performance gain compared to e-type II in ideal conditions. RAN4 to extend the feasibility analysis to higher ranks and different payload sizes to determine consistency of performance improvement. FFS on how to quantify the acceptable performance gain. (performance-complexity tradeoff)** 2. **All UE preferred encoders should be aligned (work well) with the chosen test decoder**   **Proposal 11: The following principles should guide the establishment of the alignment aspect of the feasibility analysis (steps 6-7):**   * **A well-trained UE encoder with the same or higher complexity compared to the Ref encoder that the test decoder was trained with, should have at least similar performance with the reference performance.** * **FFS on how to define a criterion for the similarity/dissimilarity of all companies' performance results with respect to the reference performance, considering UE encoder complexity as well.**   **Proposal 12: RAN4 to employ the detailed procedures shown in the flowchart for studying the feasibility analysis of test decoder Option 4a and Option 4b based on the aggregate dataset and database of test decoder models**   1. **The architecture of the test decoder is partially specified through some high-level parameters (to be agreed) while leaving room for vendor specific implementations** 2. **Interested companies implement a test decoder according to the partially specified agreement** 3. **The partially specified test decoder is trained from an aggregate data set of UE encoder latent space distributions (dataset is stored in a server). Multiple test decoders are trained from participating companies. The test decoders are stored in a database.** 4. **Each UE vendor tests its own preferred Encoder with the test decoders stored in the database** 5. **Analysis of results to determine feasibility of test decoder**   **Proposal 13: 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**   **Proposal 14: Investigate CSI similarity of input data distributions using Angular-Delay Channel Entropy (ADCE), as a statistical measure. This approach aims to reduce the reliance on more expensive monitoring procedures, such as those based on KPIs and ground truth availability, by initiating these procedures only when significant changes in data distribution are detected. The ADCE can be defined as ADCE, where , and is a virtual channel representation in angular/delay domain**  **Proposal 15: RAN4 to investigate defining metrics for Monitoring CSI Input Distribution to Trigger Monitoring Procedures and their Application to Dynamic CSI Compression**  **Proposal 16: For Monitoring CSI AI/ML Models investigate embedding a subset of known values, referred to as redundancy elements, into the CSI vector at the UE before encoding. These elements act as reference points that can be tracked throughout the encoding and decoding processes. The approach is designed to facilitate the monitoring of AI/ML models by providing a measurable and consistent benchmark for evaluating model performance.**  **Proposal 17: We propose a two-stage approach for CSI model monitoring, first distinguishing whether performance degradation stems from the AI/ML model or external factors, and then isolating the specific issue within the model if identified as the cause.**  **Proposal 18: RAN4 to investigate a performance monitoring framework to detect issues such as data drift, model degradation, and AI/ML model aging, triggering actions like re-training, re-configuration, or model switching. The system will ensure high detection accuracy, real-time monitoring, and proactive predictions, enabling reliable adaptation to changing channel conditions while maintaining optimal performance.**  **Proposal 19: For a performance monitoring framework, event detection for data distribution drift and anomalous behavior is based on the output of a trained model that predicts performance deviations (the event). Consider the following cases as shown below:**  **(1) Indirect Event Detection: AI/ML model for SGCS EST is trained to predict SGCS from latent space c. Subsequently a detection logic can trigger event detection (2) Direct Event Detection is based on AI/ML model to directly predict an event**  **Proposal 20: Consider a new KPI for CSI compression that could offer an indication of spectral efficiency loss due to misalignment of precoders. It can be conceptually described as where Vc is the corrupted reconstructed CSI**    **Proposal 21: Investigate the benefits of defining a new KPI for CSI compression that is linked to capacity loss. For an SVD receiver this metric is approximated as , where is defined as the inter-stream interference to layer k caused by the mismatch between and V. Also is the SNR and is the eigenvalue of the channel**  **For rank 1: the proposed metric becomes , while . Under ideal conditions the term due to the orthogonality. But under imperfection this term contributes to the denominator and the capacity decreases.**  **Proposal 22: Investigate the feasibility of using the spectral efficiency KPI to determine the CQI adjustment for option 1b.** |
| R4-2503773 | CMCC | ***Proposal 1: it is proposed to discuss if option 3 is feasible, whether option 4 is still needed.***  ***Proposal 2: it is proposed to discuss and allign on which part of training dataset are used for testing and/or inference .***  ***Proposal 3: it is proposed to discuss the “lifetime” of the data/model in a case by case manner. At least following cases are considered:***   * ***data/model captured in the TS*** * ***data/model uploaded as references for requirements definition*** * ***data/model uploaded for companies allignement*** |
| R4-2504006 | ZTE | ***Observation 1. We used a decoder from a company and train the encoder with our own dataset, the results without quantization and with quantization is about 0.72 and 0.66.***  ***Observation 2. From our simulation results, we use mixed dataset and five companies datasets to do test, the results of unquantization is about 0.71, and quantization is about 0.65 for mixed dataset.***  ***Observation 3. For track 2, we used a mixed dataset and our own dataset for training, and tested it with the mixed datset and our own dataset respectively. Firstly, we used the mixed dataset for training and the mixed dataset for testing. The results without quantization and with quantization are 0.61 and 0.59. We also used own dataset for training and tested it with the mixed dataset and our own dataset. The results without quantization and with quantization were 0.61, 0.63 and 0.58, 0.61 respectively.***  ***Observation 4. Option 4b is more likely option 3 if RAN4 reuse the mixed dataset as label data.***  ***Observation 5. Two different approaches are observed for option 4b, the firstly one is that DUT vendor directly uses reference model as implementation, and the second is that DUT vendor needs to develop new encoder based on the reference encoder.***  ***Observation 6. Since CSI compression is currently in the study item, we think it is better to capture these conclusion in the TR.***  ***Observation 7. The selection criteria needs to be based on practical perspective, such as complexity and performance.***  ***Proposal 1. Propose to consider a mixed dataset during training for option 4b as starting point.***  ***Proposal 2. Propose to capture the conclusion and reference model structure and parameters in TR.***  ***Proposal 3. Propose to follow RAN1’s conclusion on defining reference model structure and parameters.***  ***Proposal 4. Propose to specify reference model on RAN4 specification.*** |
| R4-2504026 | Vivo | **Proposal 1: The feasibility study of Option 3, Option 4a-1 and Option 4b could work in parallel.**  **Observation 1: For track 1, the SGCS results of Option 3 using models of Samsung, Huawei and CATT, using vivo’s dataset, are shown in Table 2.1-2 with quantization and Table 2.1-3 without quantization. The SGCS results of Option 4a/4b using Samsung model, using vivo’s dataset, are shown in Table 2.1-4 with quantization and Table 2.1-5 without quantization.**  **Observation 2: For track 1 where the training dataset is different from the dataset for reference model, it seems that Option 4b has best validation performance, Option 4a has medium validation performance, while Option 3 has worst validation performance.**  **Proposal 2: Suggested next steps for Option 4b:**   * **Case 1 (Mandatory, target RAN4#114bis): For the case that UE directly uses reference encoder in field.**   + **Take Samsung encoder and mixed dataset, train own decoder and then test own trained decoder with Samsung (reference) encoder.** * **Case 2 (Optional, maybe target RAN4#115): For the case that UE uses reference encoder to develop encoder in field.**   + **Train a new decoder using the Samsung encoder and mixed dataset. Then train an "own" encoder with the newly trained decoder.**   + **Then all companies share their "new" decoders that have been trained with the Samsung encoder and mixed dataset.**   + **Test "own" encoder vs. All other companies’ decoders from step 2. (The other companies’ decoders here represent different TE vendors.)**   **Observation 3: The SGCS results of Option 3/4a/4b using of Samsung models, are shown in Table 2.2-1 with quantization and Table 2.2-2 without quantization. The SGCS results of Option 3/4a/4b using of Nokia models, are shown in Table 2.2-3 with quantization.**  **Observation 4: For track 2 where the training dataset is the same dataset for reference model, it seems that Option 3, 4a and 4b have similar validation performance. In field, the actual training dataset (field data) will be different from the synthetic dataset of reference model**  **Proposal 3: Reference encoder needs to be specified in TS, due to the following aspects:**   * **To be used as verification encoder in the TE test decoder verification.**   + **No extra training procedure to train the verification encoder will be needed.**   + **No performance degradation from imperfectly matched verification encoder will be caused.** * **To achieve better field performance.** * **To facilitate Rel-19 AI/ML study for reference model in RAN1 and RAN4.**   **Proposal 4: Criterion for selection of reference encoder: reference encoder is selected from the encoder and decoder pair, from which the test decoder is selected.**  **Proposal 5: It is more suitable for RAN4 to specify the reference encoder, since the encoder and decoder pairs are generated and selected in RAN4.**  **Proposal 6: Reference dataset is beneficial to be specified, at least for TE test decoder verification.**  **Proposal 7: In Option 4a, the model structure needs to be specified, to guarantee the performance of test decoder.**  **Proposal 8: In Option 4a, dataset of raw channel or encoder input needs to be aligned, by aggregating the dataset of raw channel or encoder input from all companies, or choosing one dataset of raw channel or encoder input from all companies.**  **Proposal 9: In Option 4b, channel generation method and reference decoder structure need to be in the spec.**  **Proposal 10: There is no essential difference between reference model and reference dataset. Reference dataset may have some disadvantages compared to reference model, and the benefit of reference dataset over reference model is unclear.**   * **Reference dataset would cause some performance loss compared to reference model.** * **The spec impact of writing model and dataset in the spec would be similar.** * **Reference dataset needs extra step compared to reference model, which is dataset generating from the aligned reference encoder and reference decoder.** * **The storage size of reference dataset is larger than reference model.** * **Reference model is more flexible than referend dataset for future release.**   **Proposal 11: Similar feasibility conclusion through the simulation campaign of test options can be draw for both Option 3 and Option 4 (Option 4a and Option 4b).**  **Observation 5: 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.**  **Observation 6: From initial results of field test, it is observed that**   * **Directly use reference model in field (Case 2): similar performance on field data as eType II for one cell and performance loss compared to eType II is observed for another cell.** * **Finetuned encoder or decoder using field data against reference decoder/encoder (Case 2A-1, Case 2A-2): performance improved compared to directly use reference model (Case 2), but still has performance loss compared to fully trained by field data (Case1). Finetuned decoder (Case 2A-2) is better than finetuned encoder (Case 2A-1).** * **Pairing of finetuned encoder and finetuned decoder (Case 2A-3, Case 2A-4): performance loss in some cases considering mismatch on training data between UE and NW.**   **Proposal 12: To achieve better field performance, the reference model should at least include the encoder part.**  **Proposal 13: 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.** |
| R4-2504217 | Samsung | **Observation 1:** In Option 3 Track-1 (private encoder-only training), over different datasets, the SGCS performance degradation of training method-2 (encoder-only training) from training method-1 (Joint en/decoder trainings) is minor, i.e., 0.14%, 0.33%, 0.74%, 3.08% depending on different datasets.  **Proposal 1:** Based on the experience of AI-CSI compression with CNN-based neural network, **the criteria to determine “whether the selected reference decoder contains the same feature as the specific dataset”** is:   * The SGCS performance degradation of training method-2 (encoder-only training) from training method-1 (Joint en/decoder trainings) is within [1]% over a the specific dataset.   **Proposal 2:** The above criteria in Proposal 1, RAN4 can identify the dataset may contain different feature from the one used for reference decoder training.  **Proposal 3:** For Option 3 (specified reference decoder), the feasibility of Track-1 “*private encoder-only training*” can be confirmed under the following conditions:   * The selected reference decoder shall be generalized by performing the training over the dataset with enough feature contained. * The dataset used for private training should have the feature distribution aligned with the one used for reference decoder, as much as possible. * Note: private encoder-only training is referred to: fixing reference decoder, and performing training for encoder in a private company-A dataset.   **Observation 2:** In Option 3 Track-2 (encoder-only training over mixed dataset), companies provided decoders have quite similar SGCS performance, with standard deviation less than 0.01 for quantization case.  **Observation 3:** In Option 3 Track-2 (encoder-only training over mixed dataset), a minor performance degradation can be expected for (1) encoder-only training by fixing the decoder, from (1) joint encoder/decoder training.  **Observation 4:** In Option 3 Track-2 (encoder-only training over mixed dataset), by using appropriate dataset in addition to the mixed dataset, the SGCS can be further improved, at least for the following method:   * Encoder-only retaining (The encoder is initialized with random weights).   *Discussion on Test decoder Option 4*  **Observation 5:** Despitethe similarity between Option 4 in RAN4 and Direction A in RAN1, the following difference should be noted:  - In Option 4-1 (Direction A in RAN1) sharing dataset is used to derive practical encoder (to be used by UE), while in Option 4a (in RAN4) sharing dataset is used to derive test decoder (to be used by TE).  - In Option 3a-1 (Direction A in RAN1) reference encoder is used to derive the practical encoder (to be used by UE), while in Option 4b (in RAN4) reference encoder is used to derive test decoder (to be used by TE).  **Proposal 4:** Before Option 3 feasibility is confirmed or disconfirmed, the discussion on Option 4 can be deprioritized, by considering the factors including  - to avoid parallel discussion with RAN1 Direction A down-selection.  - the study on the feasibility of Option 4b can be speed up if Option 3 is confirmed. |
| R4-3504231 | Nokia | **Observation 1: Even with further alignment on simulations, the performance of the testing phase in Option 3 Track 1 is not consistent while testing is done using different companies’ datasets.**  **Observation 2: The SGCS values obtained in Option 3 Track 1 are not acceptable for several tests with different companies’ datasets, particularly with testing datasets from CATT and Qualcomm.**  **Proposal 1: Based on our results, Option 3 Track 1 does not look like a feasible option. Therefore, RAN4 should down prioritize Option 3 Track 1.**  **Observation 3: Compared to the SCGS performance from Option 3 Track 1, it is evident that Option 3 Track 2 results are more consistent among different companies’ models.**  **Proposal 2: Since the performance when using a mixed dataset for the training of both encoder and decoder is better than when using different datasets, RAN4 should pursue Track 2 over Track 1 for the feasibility study of test decoder Option 3.**  **Proposal 3: In the conclusion of the feasibility study of Option 3, and based on our results, we suggest selecting Nokia’s decoder from Track 2 as the test decoder together with the agreed mixed dataset**  **Observation 4: Using a decoder trained for the purpose of training the UE vendor’s encoder as the test decoder violates the separation of knowledge between the UE and TE vendors. This creates potential problems in the determination of feasibility of Option 4a.**  **Proposal 4: Clarify in Step 6 of the Option 4a feasibility procedure that the training of the “own decoder(s)” for the purpose of using it as a test decoder must be separated from the training of the “own encoder(s)” since the actions of different vendors (TE vendors and UE vendors, respectively) are being mimicked in this step.**   * **Step 6: Companies bring results for independent training of “own encoder(s) and decoder(s)” with the selected dataset(s)**   + **A decoder may be trained to support the training of the own encoder, but this decoder is distinct from the TE decoder used for performance assessment of Option 4a.**   + **Performance alignment to be checked/discussed**   **Observation 5: Use of the decoder in Step 5 of the Option 4b feasibility study procedure as both the TE test decoder and for training the UE encoder violates the assumption of independent training of the test decoder and UE encoder.**  **Proposal 6: Adjust Step 6 in the feasibility study procedure for Option 4b to:**   * **Step 6: Company trains “own encoder(s)”**   + **A decoder may be trained to support the training of the own encoder, but this decoder is distinct from the decoder in Step 5.**   + **The dataset used for the training is the company’s own dataset, but a dataset different from the one used in Step 5.**   **Proposal 7: For Option 3, the corresponding encoder to the selected test decoder should be the candidate for reference encoder.**  **Proposal 8: For Option 4b, encoder based approach, the selected encoder (from Option 3) should be the candidate for reference encoder.**  **Proposal 9: For Option 4a, RAN4 needs to further discuss the selection criterion for reference encoder.**  **Proposal 10: To verify test decoder and to define performance requirements, RAN4 needs to specify reference encoder. It should be captured in the TR conclusion as well as in the TS.**  **Proposal 11: If RAN1 also specifies a reference encoder, there should be alignment between RAN1 and RAN4 reference encoder.** |
| R4-2504236 | Oppo | **Proposal 1: Principles to define test decoder(s)**   * **to meet the minimum performance requirement in RAN4 tests** * **to be a simple design**   **Proposal 2: For RAN4 test model alignment, the motivation should be clarified**   * **Option1: It is expected that the alignment of minimum performance understanding can be achieved through results from different companies** * **Option2: It is expected to strictly align the output results of each company's AI/ML models through simulation**   **Proposal 3: Prioritize discussing Option 3 first and then evaluate Option 4 if needed, rather than initiating parallel discussions on both Option 3 and Option 4.**  **Observation 6: Several challenges that can prevent a seamless conversion of ONNX models to other formats, e.g., TensorFlow or PyTorch. Some main issues are listed below:**   * + - * **Lack of official conversation tools**       * **Version Compatibility Issues**       * **Conversion for custom or unsupported layers**   **Proposal 4: To reconsider the format in which the test model is delivered, such as providing it in PyTorch format, thereby relax the challenges associated with version and format conversion.** |
| R4-2504321 | Huawei, HiSilicon | ***Proposal 1:*** Use SGCS as a key metric for verifying the feasibility of procedures related to the standardization of full models.  ***Proposal 2:*** Define SGCS margin by considering the following options.   * Option 1: Average of SGCS * Option 2: SGCS @90% of CDF   ***Proposal 3:*** Define multiple degrees of performance convergence based on SGCS margin:   * SGCS margin > 0.1 indicates low - degree convergence. * SGCS margin within [0.05, 0.1] indicates medium - degree convergence. * SGCS margin < 0.05 indicates high - degree of convergence.   ***Proposal 4:*** SGCS results for Track 1 with encoder model structure under current assumption are provided in Table 1.  ***Proposal 5:*** SGCS results for Track 2 with encoder model structure under current assumption are provided in Table 1.  ***Observation 3:*** In track 2, nearly no SGCS performance loss with fixed selected decoder using mixed dataset to retrain the encoder, assuming the encoder model structure is aligned between training and retraining.  ***Observation 4:*** In track 2, around 0.05 SGCS performance loss is observed with fixed selected decoder using our own dataset to retrain the encoder.  ***Proposal 6:*** Use track 2 to move forward by both considering encoder retraining under mixed dataset and companies own dataset.  ***Proposal 7:*** SGCS results for lower-complexity CNN and transformer encoder are provided in Table 3.  ***Observation 5:*** Under the same backbone, 0.03 SGCS performance loss is observed for encoder retraining with a lower-complexity CNN-based model structure.  ***Observation 6:*** Under a different backbone from current CNN-based assumption, around 0.1 SGCS is observed for encoder retraining with a lower-complexity Transformer-based model structure.  ***Proposal 8***: Use lower complexity encoder model structure for testing/reference model structure alignment if AI-CSI turns to WI.  ***Observation 7:*** Around 0.05 SGCS performance loss can be observed in Track 1 under mixed testing dataset.  ***Observation 8:*** Around 0.1 SGCS performance loss can be observed in Track 1 under companies’ individual dataset. |
| R4-2504439 | Ericsson | [Observation 1 For track 1, there is variation in performance between different companies that depends on the training set and the test set.](#_Toc194064891)  [Observation 2 Based on the Samsung decoder, option 3 track 1 seems basically feasible, at least where companies have similar datasets.](#_Toc194064892)  [Observation 3 The variation between the mixed testset performance and individual companies testsets for track 2 has decreased but is still significant.](#_Toc194064893)  [Observation 4 Variations between different companies datasets (and the mixed dataset) appears to be related to differing levels of correlation between the SGCS observed for different sub-bands.](#_Toc194064894)  [Observation 5 Based on working with the Samsung decoder as a frozen decoder, option 3 track 2 works very well if the UE vendor encoder is trained with the same mixed dataset as the test decoder. If trained with a different dataset, there is a small difference in performance.](#_Toc194064896)  [Observation 6 For option 4, if a frozen encoder is taken from track 1, a new decoder (trained on the frozen encoder) and encoder (trained on the new decoder) can be trained that show a small performance drop compared to the encoder/decoder trained directly.](#_Toc194064897)  [Observation 7 For option 4, if a frozen encoder is taken from track 2, a new decoder (trained on the frozen encoder) and encoder (trained on the new decoder) can be trained that show a small performance drop compared to the encoder/decoder trained directly.](#_Toc194064898)  [Observation 8 For option 4b, most likely a dataset needs to be standardized to enable reproducibility of the test decoder between different TE vendors.](#_Toc194064899)  [Proposal 1 If there is a need for it to be part of a TS or TR, then preferably the reference encoder should be part of a TS. However, it may be suitable that it is stored in the same managed location as the test decoder.](#_Toc194064900)  [Proposal 2 The reference encoder should be the encoder that is paired with, i.e. trained together with, the test decoder using the same training data.](#_Toc194064901)  [Proposal 3 Keep the test decoder and reference encoder together, in the RAN4 specifications. Or alternatively store and maintain the models together in the same place and refer from the RAN4/5 specifications.](#_Toc194064902)  [Proposal 4 If available, store the dataset used to create the test decoder such that it is linked to the test decoder and reference encoder and maintained across releases.](#_Toc194064903)  [Proposal 5 Discuss during development of the requirement whether to describe test and verification data by means of a channel model or an explicit dataset.](#_Toc194064904)  [Proposal 6 The study on option 3 should consider that a 3GPP internal method for capturing test decoder AI models is needed (for CSI compression and future models). Possible means to achieve this may be discussed if time allows.](#_Toc194064905)  [Proposal 7 The study should consider how a procedure for updating the test decoder would work, considering in particular whether updates should always be backward compatible or whether signalling of versions could be needed.](#_Toc194064906)  [Proposal 8 The study should consider what key characteristics should drive the discussion on structure for a real standardization exercise. Some characteristics include:](#_Toc194064907)  [ Backbone (CNN, transformer etc.)](#_Toc194064908)  [ Complexity target](#_Toc194064909)  [ Number of models (depending on scenarios etc.)](#_Toc194064910)  [ A metric and acceptable performance target](#_Toc194064911)  [ Ability of the agreed structure for the test decoder to allow for variation of real encoder/decoder implementation](#_Toc194064912)  [Proposal 9 Conclude that, if Option 3 is supported for performance testing, then the fully specified test decoder is the same as the decoder in the reference model for Direction C in RAN1 inter-vendor training collaboration discussion.](#_Toc194064913)  [Proposal 10 Conclude that, for a real WI, a sufficient diversity of simulation assumption sets would be needed to make the training dataset for the test decoder. Also, some steps to ensure alignment of different companies’ datasets are needed.](#_Toc194064914)  [Proposal 11 To test the feasibility of option 4, all companies should upload at least a “decoder 2” (decoder trained using frozen encoder and mixed dataset) and potentially also a “decoder 3” (decoder trained using frozen encoder and company’s own dataset). This will allow interoperability between UE vendor encoders and different TE vendor decoders to be emulated.](#_Toc194064915)  [Proposal 12 Merge options 4a and 4b to standardization of a dataset and standardization of a reference encoder.](#_Toc194064916) |

## 2.2 Open issues summary

### 2.2.1 Sub-topic 2-1 Simulation analysis for options 3 and 4

This sub-topic addresses the simulation analysis and the results provided for options 3 and 4. A summary of the simulation results is provided by Huawei and available in the inbox.

In this sub-topic, some potential preliminary observations based on the results that have been presented and observations in the submitted contributions are suggested. The intention is that, during the meeting, an attempt can be made to draw up preliminary observations that are mutually agreeable and also to highlight what further work is needed to progress the conclusions.

**Issue 2-1-1: Preliminary observations on datasets**

The following are some moderator suggestions for observations about the dataset, to be discussed/refined during the meeting:

* Datasets consist of training data and separate test data.
* All companies have produced datasets based on the same simulation assumptions, however differences can be observed between performance with different companies datasets.
* The “mixed dataset” provides much more consistent results than training or testing using different individual company datasets
* The exercise demonstrates that in order to agree on test decoder(s) and or reference encoder(s), some effort is needed to align dataset(s).

**Issue 2-1-2: Preliminary observations for option 3 track 1**

The following are some moderator suggestions for discussion about option 3 track 1, to be discussed/refined during the meeting. The intention is not to draw final conclusions, but to endorse what observations can be made thus far on the results. The results numbers and conclusions will be updated when further results and analysis is available.

* It is possible for UE vendors to train an encoder based on another company’s frozen test decoder that is functional
* The performance of the encoder is very dependent on the dataset used for training the encoder and the dataset used for training the frozen decoder, and the testset
* The variation in the performance depending on the training datasets and test datasets can be as much as 30-50% in SGCS. Further details are available in the results spreadsheet.
  + Preliminary observation; further results may be added/refined and the observation updated
* For option 3, track 1 to be useful, more careful attention would be needed to the alignment of datasets
* Up to now, the same encoder structure is considered as used when creating the reference decoder; the impact of different encoder structure is not yet considered

**Issue 2-1-3: Preliminary observations for option 3 track 2**

The following are some moderator suggestions for discussion about option 3 track 2, to be discussed/refined during the meeting. The intention is not to draw final conclusions, but to endorse what observations can be made thus far on the results. The results numbers and conclusions will be updated when further results and analysis is available.

* There is reasonable SGCS convergence when companies train an encoder-decoder pair using the mixed dataset and test using the mixed testset
  + Around +- 0.5% variation in SGCS from the average; see results spreadsheet for more details
  + Preliminary observation; further results may be added
* It is possible for UE vendors to train an encoder based on a frozen test decoder that is functional
* When training of the frozen decoder, “own” encoder and testing all use the mixed dataset, there is good convergence in average SGCS results from results presented so far
  + Around +- 2% variation in SGCS from the average; see results spreadsheet for more details
  + Preliminary observation; further results may be added
* When training of frozen decoder is on the mixed dataset, “own” encoder is on a different dataset and testing is on mixed testset, results have somewhat more divergence
  + A couple of initial results show +- 3% SGCS from average, but limited data available.
* When training of frozen decoder is on the mixed dataset, “own” encoder is on a different dataset and testing is on a different dataset, results have somewhat more divergence
  + A couple of initial results show +- 3 or 4% SGCS from average, but limited data available.
* There may be some risk of overtraining the encoder
* Up to now, the same encoder structure is considered as used when creating the reference decoder by most companies; the impact of different encoder structure is not yet considered
  + A couple of initial results suggest similar SGCS can be obtained with lower complexity CNN, but SGCS would drop with a lower complexity transformer encoder

**Issue 2-1-4: Preliminary conclusions for option 4**

The following are some moderator suggestions for discussion about option 4, to be discussed/refined during the meeting. The intention is not to draw final conclusions, but to endorse what observations can be made thus far on the results. The results numbers and conclusions will be updated when further results and analysis is available.

* It is feasible to train a functional “own” decoder based on the frozen encoder, and then a functional “own” encoder based on the “own” decoder
* Results are convergent if the mixed data and test-set is used, but more divergent if other test-sets are used
  + -0.75 to +1.5% divergence in average SGCS observed from preliminary results for mixed set
* Interoperability of “own” encoder with other company “own” decoders has not yet been examined.
* There may be some risk of overtraining for the “own” decoder and “own” encoder
* Option 4a and 4b are essentially the same in this evaluation when the mixed dataset is used. One way to differentiate could be to use the mixed dataset for training the decoder in 4a and the “own” dataset in 4b
* Up to now, the same encoder structure is considered as used when creating the reference decoder; the impact of different encoder structure is not yet concluded

**Issue 2-1-5: Next step for option 3**

* Proposals
  + Option 1: For track 1, check consistence of performance across all companies testsets (Apple proposal 1)
  + For track 2, check consistency of performance using the mixed dataset (Apple proposal 3)
* Recommended WF
  + Moderator recommendation: Companies are encouraged to bring more results to fill in the blank fields in the results spreadsheet

**Issue 2-1-6: Next step for option 4**

Revised description of option 4 at RAN4#114

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

* + - Step 1-3: Reuse results of Option 3
    - Step 4: Select one or more Eigenvalue dataset(s) for further analysis based option 3
      * Selection criteria: select the dataset(s) generated from the selected encoder and decoder pair(s) from Option 3 track 1
      * Or mixed dataset from track 2
      * If multiple datasets are selected (as was the case of Option 3 track 1), the subsequent steps of the feasibility procedure will be applied on each one of the selected datasets separately
    - Step 5: Label selected dataset with encoder input/output using encoder corresponding to the selected test decoder from option 3.
    - Step 6: Companies bring results for training of “own encoder(s) and decoder(s)” with selected dataset(s)
    - Performance alignment to be checked/discussedStep 7: Conclude on overall feasibility of Option 4a
      * feasibility criteria to be discussed (e.g. Perform testing of all the UE encoders using all test decoders (replicating the process of testing UE’s against RAN4 requirements with different TE vendor decoders, others given in R4-2415376, etc ).
      * Dataset for feasibility evaluation to be discussed; may be common test dataset or own test dataset

**Option 4b (encoder based) for 2-sided model**

* Step 1-3: Reuse results of Option 3
* Step 4: Select one or more encoder(s) for further analysis based on option 3
  + Selection criteria: select the encoder(s) from the selected encoder and decoder pair(s) of Option 3 track 1
  + Or encoder(s) from the selected encoder and decoder pair(s) of Option 3 track 2
* Step 5: Company brings results for training of “own decoder(s)” with selected encoder(s)
  + Performance alignment to be checked/discussed
  + FFS what is the dataset assumed for the training
* Step 6: Company trains “own encoder(s)” with “own decoder(s)” from step 5
  + FFS what is the dataset assumed for the training
* Step 7: Conclude on overall feasibility of Option 4b
  + Feasibility criteria to be discussed (e.g. Perform testing of all the UE encoders using all test decoders (replicating the process of testing UE’s against RAN4 requirements with different TE vendor decoders, others given in R4-2415376, etc ).
  + Dataset for feasibility evaluation to be discussed; may be common test dataset or own test dataset

Encoder in step 5 should at least have the agreed structure. Companies can also bring analysis/results with other encoders (using different structures)

* Proposals
  + Option 1: Companies test their “own new encoders” with other companies “new decoders” (New encoders/decoders trained using Samsung frozen encoder) (Vivo proposal 2, Ericsson proposal 11)
  + Clarifications for step 6 (Nokia proposals 5-6)
* **Option 4a Step 6: Companies bring results for independent training of “own encoder(s) and decoder(s)” with the selected dataset(s)**
  + **A decoder may be trained to support the training of the own encoder, but this decoder is distinct from the TE decoder used for performance assessment of Option 4a.**
  + **Performance alignment to be checked/discussed**
* **Option 4b Step 6: Company trains “own encoder(s)”**
  + **A decoder may be trained to support the training of the own encoder, but this decoder is distinct from the decoder in Step 5.**
  + **The dataset used for the training is the company’s own dataset, but a dataset different from the one used in Step 5.**
  + Option 2: Merge option 4a and 4b (Ericsson proposal 12)
* Recommended WF
  + Moderator recommendation:
    - Agree updates to step 6
    - Companies are encouraged to continue to bring results to fill in the blanks in the spreadsheet
    - As many companies as possible upload “own” decoder based on Samsung frozen encoder within 1 week
    - Companies provide results operating their “own” encoder with other companies “own” decoders to check interoperability

**Issue 2-1-7: Clarifications**

* Proposals
  + Option 1: Use mixed dataset for option 4b (ZTE proposal 1)
  + No need to work further on option 4 if option 3 feasible (CMCC proposal 1)
  + Work on option 3 and 4 in parallel (Vivo proposal 1)
  + Work on option 3 first, then option 4 (Samsung proposal 4, Oppo proposal 3)
  + Deprioritize option 3 track 1 (Nokia proposal 1, 2, Huawei proposal 6)
  + Use Nokia for test decoder for track 2 (Nokia proposal 3)
  + Whether the test model alignment is to align minimum performance or strictly align outputs (Oppo proposal 2)
  + Use PyTorch instead of ONNX (Oppo proposal 4)
  + The encoder/decoders considered during the SI will not be usable for the WI stage (NTU proposal 3)
* Recommended WF
  + Moderator recommendations:
    - The results spreadsheet contains columns for mixed and own dataset training, corresponding to the proposals for options 4a and 4b. Companies can contribute into this spreadsheet to consider the difference.
    - Do not discuss prioritization. Companies encouraged to contribute as much data as possible in order to be able to make conclusions by August.
    - Agree that it is clear that the encoders/decoders uploaded during the SI phase will not be useful for a WI.

**Issue 2-1-8: New proposal – Option 4c**

* Proposals
  + Add an option 4c, in which an encoder-decoder pair is evaluated based on latent message re-ordering considering the quantized dominant Eigenvectors (NTU proposals 5-7)
* Recommended WF
  + TBA

**Issue 2-1-9: Metric for evaluating inference performance with options 3/4**

* Proposals
  + Use SGCS as key metric (Huawei proposal 1)
  + Further discuss whether to consider average or 90% CDF of SGCS (Huawei proposal 2)
  + Criteria for deciding on alignment (Huawei proposal 3):
  + SGCS margin > 0.1 indicates low - degree convergence.
  + SGCS margin within [0.05, 0.1] indicates medium - degree convergence.
  + SGCS margin < 0.05 indicates high - degree of convergence.
  + 1% SGCS difference between joint trained encoder/decoder and separately trained encoder (Samsung proposal 1-3)
* Recommended WF
  + Average SGCS is already provided by many companies
  + For the criteria of procedure feasibility, consider the following options.
    - Option 1: Use the variation in SGCS from the average among companies.
      * FFS: Absolute value (SGCS difference span) or percentage (SGCS difference span over SGCS average)
      * FFS: Whether to separately consider the variation in SGCS from the joint training
  + Option 2: Use 1% difference between jointly trained performance and separately trained performance
  + Regarding the criteria, probably not enough time to agree in the meeting, but discuss further in a later meeting when more results are available.

**Issue 2-1-10: Metric for comparing datasets**

* Proposals
  + Option 1: Angular-Delay Channel Entropy (ADCE) to compare training datasets (Apple proposal 2)
* Recommended WF
  + TBA

**Issue 2-1-11: Further investigations**

* Proposals
  + Investigate performance loss when one or both parts of the model are specified with synthetic dataset (Apple proposal 4)
  + Investigate what happens if different structure is used for the UE encoder (Apple proposal 7, 11)
  + Investigate 2 layers performance (Apple proposal 8)
  + Performance gain from test decoder (Apple proposal 10)
  + What level of similarity/difference between different companies performance results is acceptable (Apple proposal 11)
* Recommended WF
  + Companies are encouraged to provide more data on difference between performance with modelled data and real world performance, as Vivo has done
  + Companies are encouraged to provide more data on the impact of using a different encoder structure, as Huawei and CATT have done.
  + Regarding 2-layers performance, need to discuss how it relates to SGCS.
  + Regarding performance gain from test decoder, do not discuss since the decoder structure has been chosen arbitrarily, not to maximize performance. Note that it would need intense discussion for a work item though.

**Issue 2-1-12: Clarification of NTU questions**

A number of clarification questions were posed by NTU. The moderator understanding, based on previous agreements, is provided below. It is suggested to further clarify offline as needed.

For option 4a

* Whether option 3 track 1 has an encoder selected besides the selected fixed decoder

[Moderator understanding] The encoder that is paired with the selected frozen decoder of option 3 can be taken for option 4a.

* If we use the encoder/decoder selected from option 3 track 1, what’s the additional alignment needed given that we agreed to have “*Each company is encouraged to train own encoder using the fixed decoder and their own dataset, and then test with every companies datasets, in order to check for consistency in results*” procedure in option 3 discussion, which is the same as step 6 and 7. Similar issue is observed if we use dataset from option 3 track 2

[Moderator understanding] The difference is that for option 3, each company trains an encoder against the selected frozen decoder. For option 4a, each company trains a new decoder based on the frozen encoder and then trains a new encoder based on their frozen decoder.

* If the mixed dataset from track 2 is used, how to train the encoder and decoder based on it given that it only contains encoder input?

[Moderator understanding] For option 3 track 2, a frozen decoder is used and an encoder is trained using the frozen decoder and the dataset. For option 4a, a new decoder (and associated encoder) is trained based on the mixed dataset and the frozen encoder. Then, afterwards there should be a check if the new associated encoders are interoperable with other companies new decoders.

* How to verify whether the following agreed aspects of option 4 are feasible or not?
  + *Test repeatability should be ensured (variation among TE vendor implementations should be bound)*
  + *Other vendors should also be able to develop such a decoder and which can deliver similar performance*

[Moderator understanding] This is achieved by means of performing inference using the “new” encoder (trained using own “new” decoder) with other companies “new” decoders. This emulates being a UE vendors and testing with different TE vendors (represented by other companies)

For option 4b

* How to select one encoder from companies’ own encoder in option 3 track 1?

[Moderator understanding] During RAN4#113, Samsung, Huawei and CATT were identified for examples of frozen encoders. After RAN4#114 (in which there was some alignment effort), these companies uploaded decoders and encoders that can be used as frozen decoder (option 3 track 1) or frozen encoder (option 4).

* Given that option 4b aims to select one reference encoder, why we need to check multiple encoder alignment in step 6 after one is selected already?

[Moderator] 3 companies were selected to add robustness to the evaluation, but in reality only one frozen encoder needs to be selected.

* How to verify whether the following agreed aspects of option 4 are feasible or not?
  + *Test repeatability should be ensured (variation among TE vendor implementations should be bound)*
  + *Other vendors should also be able to develop such a decoder and which can deliver similar performance*
* [Moderator understanding] This is achieved by means of performing inference using the “new” encoder (trained using own “new” decoder) with other companies “new” decoders. This emulates being a UE vendors and testing with different TE vendors (represented by other companies)

### 2.2.2 Sub-topic 2-2 Other considerations

In this section, considerations that do not directly relate to the simulation campaign are discussed:

**Issue 2-2-1: What needs to be specified for option 3**

During RAN4#114, the following agreement on reference model was made:

By definition, a test decoder is standardized for option 3. The need to define a reference encoder was discussed and the following agreement reached

Agreement:

* To verify the performance of CSI compression test decoder for calibration of test equipment, agree to define the reference encoder including encoder structure and parameters in RAN4
  + FFS on whether to specify it in TS or TR, and capture the conclusion in the TR conclusion part
  + FFS on criterion for selection of reference encoder
  + FFS on decision whether RAN4 or RAN5 will specify it
* Proposals
  + For the reference encoder (already agreed) take the encoder corresponding to the test decoder (CATT proposal 2, Vivo proposal 4, Nokia proposal 7, Ericsson proposal 2)
  + FFS where to capture reference encoder (CATT proposal 3)
  + Specify the reference encoder in the TS (Vivo proposal 3, Ericsson proposal 1)
  + Specify the reference encoder in TR and TS (Nokia proposal 10)
  + Specify the reference encoder in RAN4 (Vivo proposal 5)
  + Store the reference encoder together with the decoder and refer to it from TS/TR as appropriate (Ericsson proposal 1, 3)
  + Training dataset (CATT proposal 4, Vivo proposal 6, Ericsson proposal 4 if available)
  + Either test and verification datasets or test and verification channel models (Ericsson proposal 5)
  + Do not specify option 4 if option 3 is feasible (CMCC proposal 1)
  + Test decoder should meet the minimum performance requirement in RAN4 and be simple (Oppo proposal 1)
* Recommended WF
  + Agree that the reference encoder will be the encoder that corresponds to the frozen test decoder.
  + Check which of the following is agreeable:
    - Store encoder and decoder models together.
    - Specify/reference the reference encoder from RAN4 TS
    - Specify/reference the reference encoder from RAN4 TR
  + Check whether it is agreeable to assume that the training dataset for the frozen decoder is stored, if available
  + FFS whether test/verification is based on stored datasets or a channel model.

**Issue 2-2-2: What needs to be specified for option 4**

* Proposals
  + Specify training dataset (Qualcomm proposal 1, vivo proposal 8)
  + Specify reference encoder (Qualcomm proposal 1, vivo proposal 10, 12)
    - For option 4b, reference encoder should be selected encoder. For option 4b FFS (Nokia proposals 8-9)
  + Specify reference decoder (Qualcomm proposal 2)
  + Specify a means to verify the test decoder performance based on the difference metric (See NTU proposal 4)
  + Partially specify test decoder structure (Apple proposal 12)
  + Fully specify model structure (Vivo proposals 7, 9)
* Recommended WF
  + Dataset is captured for option 4a. Check whether it is agreeable that dataset is stored for option 4b.
  + Check if agreeable to assume reference decoder is specified.
  + Check whether agreeable that test decoder is (i) fully specified or (ii) partially specified
  + Discuss if verification of the decoder is based on verification dataset, channel model or the proposal from NTU

**Issue 2-2-3: Difference of option 3 and RAN1 inter-vendor collaboration direction C**

* Proposals
  + Option 1: Clarify whether the RAN1 and RAN4 discussions differ (Apple proposal 5)
  + Option 2: Capture reference model in RAN4 specificaiton (ZTE proposal 4)
  + Option 3: Align RAN4 and RAN1 reference encoders (Nokia proposal 11)
  + Option 4: Option 3 in RAN4 and direction C in RAN1 is the same (Ericsson proposal 9)
* Recommended WF
  + Check if agreeable that RAN4 option 3 and RAN1 direction C are the same
  + An eventual WI would then need to decide whether the decoder/encoder is captured in RAN1/RAN4 or both.

**Issue 2-2-4: Number of test decoders to be specified for option 3**

* Proposals
  + Option 1: Discuss how many test decoders to be specified for option 3 (Apple proposal 6)
* Recommended WF
  + Wait for further progress and discuss in a future meeting

**Issue 2-2-5: Considerations on dataset/model capturing**

* Proposals
  + Option 1: Discuss and align on which parts of datasets are for training and which for testing (CMCC proposal 2)
  + Discuss the lifetime of data and models considering TS, references for requirements and use for company alignment separately (CMCC proposal 3)
  + 3GPP will need to develop a means for capturing models (Ericsson proposal 6)
  + There is a need to consider a procedure how models would be updated (Ericsson proposal 7)
* Recommended WF
  + Wait for further progress and discuss in a future meeting

**Issue 2-2-6: How to decide on reference model structure**

* Proposals
  + Option 1: Follow RAN1 conclusion (ZTE proposal 3)
  + Option 2: Use lower complexity model structure for a WI (Huawei proposal 8)
  + Option 3: Consider at least complexity backbone, complexity target, number of models, performance target, the ability to allow for variation of real encoder / decoder (Ericsson proposal 8)
* Recommended WF
  + TBA

**Issue 2-2-7: What should be in the dataset used for creating RAN4 models**

* Proposals
  + Option 1: TDL, CDL and Uma mixture (Vivo proposal 13)
  + Option 2: A diversity of simulation assumption sets needed. Need to ensure alignment if different companies contribute (Ericsson proposal 10)
* Recommended WF
  + TBA