**3GPP TSG-RAN WG4 Meeting # 116 R4-25XXXXX**

**Bengaluru, India, 25th ‒ 29th August, 2025**

**Agenda item:** 7.18.1

**Source:** Moderator (Ericsson)

**Title:** Topic summary for [116][132]NR\_AIML\_air\_part2

**Document for:** Information

# Introduction

This is the summary for the Study Item on 2-sided CSI compression. In addition, the general agenda item 17.2 for the AIML PHY work item is included in the summary, since the contributions for 17.2 (AI/ML PHY WI general) and AI 18.2 (AIML PHY SI general) overlap.

# Topic #1: General considerations

In this topic, general considerations (AI 17.2 and 18.2) are considered.

For Life Cycle Management, there are a number of proposals that are specific to LCM for Beam Management, and a number of proposals that are specific to LCM for CSI prediction. These are handled separately in issues 1-1 and 1-2. If conclusions can be reached on LCM for these two use cases, issue 1-3 and 1-4 will aim to check whether any general conclusions can be made applicable for other use cases and the 2-sided study.

In addition, post-deployment testing, generalization and principles for setting requirements are discussed.

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| R4-2510879 | Huawei | Spreadsheet containing model structures and simulations from previous meetings |
| R4-2510880 | Huawei | Simulation results collection |
| R4-2509429 | Apple | 1. RAN4 is invited to discuss a composite performance metric for AI/ML-based functionalities that jointly captures both prediction accuracy (e.g., throughput gain or beam selection accuracy) and system efficiency (e.g., reduction in measurement or reporting overhead). Such a composite metric would better reflect the true benefits of AI/ML approaches by acknowledging trade-offs between slight degradations in accuracy and significant gains in control resource savings 2. RAN4 to study the applicability of the existing LCM framework defined for UE-sided AI/ML models to the UE component of two-sided AI/ML models. This includes evaluating whether the activation, deactivation, switching, and fallback procedures defined for single-ended models can be reused without modification, or whether additional signalling and synchronization are needed between UE and gNB to maintain consistent inference behaviour 3. RAN4 to define performance monitoring requirements for two-sided CSI compression use cases, separately for UE-side monitoring and NW-side monitoring. For NW-side monitoring, legacy delay requirements may be reused, as the UE reports CSI via legacy codebooks. For UE-side monitoring, new delay requirements should be studied, reflecting the non-legacy nature of the reported metrics (e.g., intermediate KPIs such as SGCS or reconstruction error). The study should also identify applicable reporting modes (periodic, semi-persistent, or aperiodic) and ensure that the specified delay budgets support timely activation or fallback decisions based on monitoring results |
| R4-2509930 | Nokia | Proposal 1: RAN4 to consider leveraging the LCM requirements being defined for the UE-sided use cases and using the Rel-19 NR-AI/ML-Air LCM conclusions as a baseline in the LCM discussions for the 2-sided models.  Proposal 2: In case of NW-side performance monitoring with target CSI reporting in the two-sided CSI compression use case, no new requirements are needed for performance monitoring reporting.  Proposal 3: In case of UE-side performance monitoring in the two-sided CSI compression use case, RAN4 to use input from RAN1 regarding the performance monitoring metric(s) to be reported by the UE to decide whether to define accuracy and delay requirements for performance monitoring reporting.  Proposal 4: RAN4 to consider the following parameters for the generalization testing in the two-sided CSI compression use case:  • Deployment scenarios  • Indoor/outdoor UE distributions  • Carrier frequencies  • TxRU mappings  Proposal 5: RAN4 to discuss the need for scalability testing in the two-sided CSI compression use case (e.g., in terms of CSI payload sizes, bandwidths and subband sizes, and Tx port numbers).  Proposal 6: RAN4 to clarify whether the agreement about post-deployment testing Option 2 prioritization (post-deployment post-activation functionality testing based on performance monitoring) is valid for the two-sided use cases also. |
| R4-2510136 | Korea Testing Laboratory | [Validation and time reference]  Observation 1: Identity mechanisms (Alt.1-Alt.4) do not reveal when a new model or functionality becomes effective; without a clear time-reference, deterministic boundary classification is not guaranteed.  Observation 2: The “effective-from” uplink indication enables Option-2 post-deployment validation using the existing performance-monitoring framework without introducing new metrics, anchors, or procedures, and with no changes to the reporting scheme; classification remains strictly on the decision timeline.  Observation 3: Multiple functions (use cases) can change close together or at different times; anchors may differ by function.  Observation 4: Per-function indications are allowed so each use case maintains a clean boundary on its own anchors (identifier values are FFS).  Observation 5: Arrival timing: the indication should arrive no later than the earliest uplink that carries any observation produced under the updated behavior; co-carrying it with that first observation is acceptable. If there is no immediate uplink opportunity (e.g., DRX, RRC Inactive), send it at the next available uplink opportunity, preferably co-carried with the first updated observation.  Observation 6: Boundary classification is anchored to the decision timeline; delivery-time variation does not reassign samples.  Observation 7: Late/duplicate/missing cases are handled minimally (late: unclassified until indication is received; duplicate: idempotent; missing: no validation window, routine monitoring continues; retry/recovery FFS).  Observation 8: Mobility robustness: mobility events (e.g., handover) do not affect classification because the boundary is anchored to decision time, not delivery time.  Observation 9: Tie-time rule: if the effective-from moment and an observation’s decision time are equal, treat the observation as post-change.  Proposal 1: Discuss adopting a minimal, explicit “effective-from” uplink indication that marks when the updated behavior applies and identifies the relevant functionality (ref. RAN1-agreed use-case/anchor group; exact identifier values are FFS).  Proposal 2: Discuss minimal delivery and classification rules to ensure deterministic boundaries.  [Distinction between validation and runtime monitoring]  Observation 10: It is preferred to treat validation and runtime monitoring as serving different purposes using distinct configuration values rather than a single unified design.  Observation 11: The distinct configuration values may include: phase, metric type, per-phase thresholds, reporting direction, and the LCM action-suppression flag; day-to-day operation (normal operation) and algorithms remain implementation-specific.  Observation 12: It is preferred that validation be a short, bounded window using fine-grained metrics and a temporary LCM action-suppression flag to assess the effect of a fine-tune or update.  Observation 13: For runtime (normal operation), it is preferred to ensure efficient, low-overhead behavior using compact, RAN1-agreed metrics (e.g., RS-PAI/SGCS) with sparse, directional reporting.  Proposal 3: Discuss a small, distinct set of configuration values for post-deployment evaluation and leave operational handling to implementation.  Proposal 4: Discuss realization of a phase indicator (e.g., validation vs runtime) and validation efficiency within the existing RAN1 performance monitoring framework. |
| R4-2510165 | CMCC | Observation 1: According to lastest RAN1 feature list, there are many feature group, and for each feature group, there are multiple components.  Proposal 1: for RAN4 previous agreements “define one test per UE capability as a minimum”, it is necessary to align the granularity of UE capability. Whether it refer to feature group or refer to each component of each feature group in RAN1 feature list.  Proposal 2: it is proposed to consider >1 test per UE capability if the granularity of UE capability is rough. |
| R4-2510241 | CAICT, Ericsson, Qualcomm, APPLE, Huawei, Hisilicon, OPPO, CATT, CMCC, NTT DOCOMO, INC., Vivo, NTU, Nokia, Xiaomi, Mediatek, Rohde & Schwarz, Samsung, Intel, ZTE Corporation, Sanechips, Korea Testing Laboratory | TP capturing the conclusions from the SI and WI into the TR |
| R4-2510341 | vivo | Observation 1: For Option B, the determination of whether UE functionality should be considered as: beginning activation, or already activated after UE receives RRCReconfiguration in Step 5 (or UE reports applicable functionalities via RRCReconfigurationComplete in Step 4) is related to RAN2’s reply to the LS about periodic CSI reporting for Option A.  Proposal 1：For the periodic CSI reporting, the activation delay requirements created is defined as:  Upon reporting applicable functionalities via RRCReconfigurationComplete in step 4 (for Option A) or upon the reception of CSI-ReportConfig for inference configuration in RRCReconfiguration message of step 5 (for Option B) in slot n, the UE shall be capable to transmit valid CSI inference report and apply actions for the functionality being activated no later than in slot n+(T\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length) , where:  Tactivation\_time is the activation delay of the functionality in milliseconds, including preprocessing for inference (e.g., loading required information for inference).  TCSI\_reporting is the delay (in ms) including uncertainty in acquiring the first available downlink CSI reference resource, UE processing time for CSI reporting and uncertainty in acquiring the first available CSI reporting resources.  Note: The definition of the above delay is based on the working assumption that "activation begins only after the UE has reported applicability reporting and the network has sent the complete inference configuration."  activation delay is defined as below:  Upon reiceiving MAC-CE triggering activation command in slot n, the UE shall be capable to transmit valid CSI inference report and apply actions for the functionality being activated no later than in slot n+(〖T\_HARQ+[3]ms+T〗\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length) , where:  Tactivation\_time is the functionality activation delay in milliseconds, including preprocessing for inference (e.g., loading required information for inference).  TCSI\_reporting is the delay (in ms) including uncertainty in acquiring the first available downlink CSI reference resource, UE processing time for CSI reporting and uncertainty in acquiring the first available CSI reporting resources.  Proposal 3: For DCI-based semi-persistent CSI reporting, except for the starting point, the activation delay definition is similar to that of periodic CSI reporting  Proposal 4: For CSI prediction, activation delay requirements should be introduced for periodic, semi-persistent, and aperiodic CSI reporting.  Proposal 5: For CSI prediction and beam management, RAN4 should not define deactivation delay requirements.  Observation 2: For BM-Case 2, the potential RRM requirement impacts mainly include the following aspects. Regarding these requirements, RAN4 has mainly focused on discussions for BM-Case 1 in the WI, while BM-Case 2 has not been thoroughly studied or discussed. Completing the definition of these requirements for BM-Case 2 within a single meeting will be a highly challenging task.  Aspect 1: Prediction delay requirement for BM-Case 2  Aspect 2: TCI state-related (including TCI state known condition and MAC-based TCI state switch delay)  Aspect 3: Metrics/KPIs for BM-Case 2 and corresponding performance evaluation  Aspect 4: Performance monitoring-related (including delay requirement and accuracy requirement)  Proposal 6: RAN4 needs to discuss whether and how to define the RRM requirements for BM-Case 2.  Alt 1: If the definition of all core part-related requirements for BM-Case 2 cannot be completed in the RAN4#116 meeting, then no requirements related to BM-Case 2 (including core part and performance part) will be introduced in Rel-19.  Alt 2: RAN4 selectively defines only some necessary requirements in RAN4#116 meeting. |
| R4-2510877 | 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. |
| R4-2509244 | NTU | **Proposal 1: In order to discuss the necessity and feasibility of the LCM decision accuracy requirement, RAN4 needs to study the following**   * **The definition of metrics to evaluate the accuracy of LCM decision: since an explicit definition of a correct LCM decision is complicated, we suggest developing the accuracy requirement based on the following principle**   **After the LCM decision, the performance metrics (can leverage performance metrics of the legacy procedures, e.g., throughput, BLER) is improved by *x*%**   * **The proper environment condition or configuration change to verify the LCM procedure and decision: RAN4 needs to study what types of changes can trigger LCM decisions (and incur performance degradation if no LCM decision is taken) so that we can design a test to verify the LCM decision accuracy.** |
| R4-2509243 | Apple | Observation 1 Option 1 (pre-activation testing) ensures assurance but can be slow, costly, and impractical for frequent updates. Option 2 (post-activation LCM monitoring) is scalable but risks false triggers, overhead, and delayed detection if not carefully designed  Observation 2 Pre-activation protects the network from immediate catastrophic failures at the moment of deployment, while post-activation protects against gradual or unforeseen degradation during the model’s operational life. Both are needed in the hybrid approach to cover different risks and time horizons  Observation 3 Shadow inference is essential for Option 2 because post-activation monitoring only gathers KPIs from the model actively serving inference. Without shadow mode, standby models remain untested in live conditions, risking undetected drift or degradation until swapped  Proposal 1 For post-deployment AI/ML operation, RAN4 shall adopt a hybrid life-cycle management (LCM) framework that combines pre-activation validation for major updates with post-activation monitoring for all updates. Major updates, defined as architecture changes, new RF front-end profiles, or significant dataset distribution shifts, shall undergo pre-activation validation in lab or on-device test mode to establish or update their KPI envelope. Minor updates may bypass pre-activation validation but must retain the previously established KPI envelope through online monitoring.  Proposal 2 For multi-model deployments under Option 2 LCM, post-activation monitoring shall include shadow inference for all non-active models. These standby models shall be exercised on the same input stream as the active model, with their outputs evaluated against genie references or active-model KPIs. This ensures drift or degradation is detected before a model is swapped into service, without impacting live operation  Proposal 3 RAN4 shall define standardized criteria for classifying model updates as major or minor within a given deployment scenario. These criteria shall consider the scope of model retraining, changes to architecture, inclusion of new channel models or RF profiles, dataset distribution shifts, and modifications to the inference pipeline. This classification will determine whether a model must pass through the pre-activation validation stage before deployment, while all models, regardless of classification, will be subject to post-activation LCM monitoring  Proposal 4 RAN4 shall specify that KPI envelopes used for post-deployment monitoring be scenario-specific and model-specific, ensuring that performance tracking is tailored to the environment and use case the model was designed for. KPI envelopes shall be derived from pre-activation validation results for major updates or inherited from prior deployments for minor updates.  Proposal 5 (Use-Case-Aware Delay Classes for Activation and Fallback) RAN4 should define a set of delay classes that reflect the time-criticality of various AI/ML functionalities (e.g., Class 1: for beam mobility; Class 2: for medium-latency AI tasks; Class 3: for background tasks). Each AI/ML Function Group (FG) or feature, once agreed in RAN1, would be mapped to one of these classes. The class assignment would be signaled during capability exchange or RRC reconfiguration, allowing the gNB and UE to align their expectations.  Proposal 6 RAN4 should define a Transition Window for activation, deactivation, and fallback of AI/ML models, starting when the UE receives the relevant command and ending when it is ready for inference or reconfigured to legacy operation. By default, control is RRC-based, with the start at reception of the RRC command and the end at the UE’s RRC Complete message, providing an unambiguous completion signal. If MAC CE/DCI-based control is used, the start is reception of the MAC CE/DCI, but an explicit UE-to-NW readiness indication, via MAC CE acknowledgment, implicit CQI feedback, or explicit signaling with the model ID, must be added to define the end point. This delay should be standardized and tested independently of the UE’s internal steps. Another option is described in Proposal 9 for BM.  Proposal 7 When a UE determines an AI/ML model is no longer applicable or its runtime KPIs fall outside the validated KPI envelope, it shall signal this state change to the network using minimal applicability/inapplicability reporting. This ensures the gNB has visibility into UE operational state, enabling coordinated actions such as adapter switching, model reconfiguration, or fallback to legacy operation.  Proposal 8 (Define Activation and Fallback Delay via First Valid Measurement Trigger) RAN4 should define AI/ML activation and fallback delay in terms of the time to the first valid L1/L3 measurement report generated under the new operational state, rather than relying on internal UE processing timelines. For example, in beam management, activation delay would be measured from the network’s activation command to the first AI/ML-based report, and fallback delay from the fallback trigger to the first legacy (non-AI/ML) report. Using an externally observable event as the endpoint removes ambiguity, enables consistent measurement across UEs and scenarios, and supports interoperability and conformance testing in line with RAN2’s applicability and activation framework  Proposal 9 For the UE-side AI/ML model in the beam management use case, RAN4 should define activation delay requirements for periodic CSI reporting, measured from the applicability report (e.g., in RRCReconfigurationComplete) to the first valid AI/ML-based report, ensuring timely operation after functionality is deemed applicable. Similarly, deactivation delay requirements should be defined, measured from the network’s deactivation command to the end of AI/ML-based reporting, ensuring prompt transition to non-AI/ML operation when a functionality becomes non-applicable. For fallback cases, the fallback delay should be defined separately if it includes additional components beyond deactivation, for example, the time until the first valid non-AI/ML report is received. |
| R4-2509760 | Xiaomi | Proposal 1: UE doesn’t need to finish the activation before sending RRCReconfigurationComplete.  Observation 1: In legacy, no de-activation delay is defined for CSI reporting in RAN4.  Observation 2: Without considering switching or fallback procedures, defining deactivation delay in isolation may lack practical relevance.  Proposal 2: For AI BM, legacy CSI-ReportConfig framework will be used, don’t need to define deactivation delay.  Proposal 3: RAN4 not to define performance monitoring delay for BM-Case1 and BM-Case2.  Proposal 4: Beam prediction accuracy is defined as the ratio of the number of correctly predicted beam instances to the total number of evaluated monitoring occasions over a configured time window, i.e. Np/N.  Where,  - N is the number of latest CSI-RS or SSB transmission occasions used for performance evaluation, as configured by the network.  - Np is the number of those occasions for which the beam prediction is considered accurate according to the matching rule (e.g., at least one of the Top-K predicted beams matches one of the Top-M strongest beams measured based on L1-RSRP).The values of K, M, and N are network-configurable parameters.  Proposal 5: For beam ID prediction accuracy, there is no performance requirement in legacy. Set a specific requirement for AI.  Proposal 6: For RSRP prediction accuracy, measurement error may have impact on whether AI requirements can exceed “legacy” requirements. |
| R4-2509920 | Nokia | Proposal 1: Based on the current RAN4 agreements for rel. 19 on performance monitoring, RAN4 to consider the test-active functionality/ configuration for post-deployment validation.  Observation 1: A test-active configuration can be provided/signalled and triggered by the Network (gNB/LMF) to the UE using RRC/MAC signalling for transitions from inapplicability to applicability or first applicability report of a non-conformance tested configuration (if detectable).  Observation 2: The newly tested AI/ML configuration (which was in test-active mode) may not necessarily be activated immediately after the testing. The actual activation decision criteria is left for the NW implementation.  Proposal 2: Based on the current RAN1 and RAN2 agreements for Release 19 on performance monitoring, RAN4 to not consider new requirements and tests to ensure the required functioning of performance monitoring procedures for the test-active functionality/ configuration.  Proposal 3: RAN4 to investigate the potential solutions which can enable improved in-field post-deployment testing/monitoring of multiple UEs resulting in efficient usage of the network and UE resources.  Proposal 4: RAN4 to define accuracy requirements for UE-assisted monitoring where the UE reports the monitoring metric (i.e., monitoring Type 1 Option 2) in BM-Case1 and BM-Case2. The accuracy of the performance monitoring metric should be tested at least in static radio conditions.  Proposal 5: RAN4 to define delay requirements for UE-assisted monitoring where the UE reports the performance monitoring metric, RS-PAI, in BM-Case1 and BM-Case2. The starting point of this delay can be when UE sends RRCReconfigurationComplete message in response to the configuration of monitoring RS resources via RRCReconfiguration, and the ending point can be when UE reports the first performance monitoring metric, RS-PAI.  Observation 3: Threshold-based criteria with respect to the reported SGCS under static channel conditions can ensure that the reported SGCS metrics reliably reflect the prediction quality of the AI/ML model, supporting consistent and interpretable performance monitoring.  Proposal 6: RAN4 to define accuracy requirements for CSI prediction performance monitoring, specifically for UE-reported SGCS1 values.  Proposal 7: RAN4 to define the delay requirement for the first performance monitoring report in CSI prediction, based on the time from RRC (re)configuration Complete to the transmission of the initial SGCS-based report, in alignment with the beam management framework.  Proposal 8: RAN4 to consider RRCReconfigurationComplete message containing applicable functionality report as the starting point to define activation delay requirement for periodic CSI reporting for AI/ML BM use case.  Proposal 9: RAN4 to consider first inference report as the end point to define activation delay requirements for periodic/aperiodic/semi-persistent CSI reporting for AI/ML BM use cases.  Proposal 10: RAN4 to not define any deactivation delay requirements for AI/ML configurations  Observation 4: For Release 19, all the NR\_AI/ML\_air feature groups (58), as currently agreed in R1-2504673, are all marked as ‘Optional with capability signalling’.  Observation 5: For Release 19, all the NR\_AI/ML\_air feature groups for beam management and CSI prediction use cases (58-0, 58-1 and 58-3) as currently agreed in R1-2504673, are all marked as ‘Need for the gNB to know if the feature is supported’  Proposal 11: RAN4 requirements for generalization must be studied at the granularity of the defined NR\_AI/ML\_air Feature Groups and, including one or more of their defined components and parameters.  Proposal 12: RAN4 should not create a mixed dataset for generalization testing since the objective is not to test the AI/ML model. |
| R4-2510104 | Mediatek | Proposal 1: For periodic CSI reporting, RAN4 can start the discussion on how to define model activation delay requirements for Option B at first.  Proposal 2: For periodic CSI reporting, UE’s activation procedure starts from RRC complete message after Step 5 for Option B.  Proposal 3: For Option B, the overall activation delay of periodic report can be divided into 2 terms: Tactivation\_time+ TCSI\_Reporting. Tactivation\_time is the duration UE needs to activate the AI functionality. TCSI\_Reporting is the duration between the time that UE activated the functionality and the time UE transmit the first CSI report in UL.  Observation 1: The Tactivation\_time is not a constant value. It depends on the initial location where the AI model is stored in UE upon receiving Step 5 RRC message as well as the model size and type.  Observation 2: It is important for NW to know when the functionality at UE side has been activated.  Proposal 4: Send LS to RAN2 to allow UE to report the time needed for functionality activation to network through RRCReconfigurationComplete message.  Proposal 5: For semi-persistent and aperiodic report, the applicable functionality shall be partially prepared before the triggering command (MAC CE/DCI).  Proposal 6: For semi-persistent report and aperiodic report, the overall delay can be partitioned into following components: Tactivation\_time1+ Tactivation\_time2+ TCSI\_Reporting. Tactivation\_time1 is the duration UE needs to partially prepare the AI functionality before UE is able to receive the MAC CE triggering. Tactivation\_time2 is the duration UE needs to fully activate the AI functionality, starting from the time UE receives the MAC CE. TCSI\_Reporting is the duration between the time that UE fully activates the functionality and the time UE transmits the first CSI report in UL.  Proposal 7: For MAC CE triggered semi-persistent report, Tactivation\_time2 = THARQ +3ms + X, where 3ms is the time for MAC CE decoding, and X is up to UE capability. If proposal 6 can be agreed, candidate value of X can be 2ms for AI/ML BM.  Proposal 8: No interruptions are expected during functionality deactivation procedure. Open to define functionality deactivation delay requirements or not. |
| R4-2510163 | CMCC | Proposal 1: for beam prediciton, for both semi-persistent CSI reporting and aperiodic CSI reporting, it is proposed that the ending point of activation delay is sending the inference report.  Proposal 2: for beam prediction, for the case that the reporting is triggered by MAC CE, the funtionality activation delay is THARQ + 3ms +Tmeasurement period for inference.  Proposal 3: for beam prediction, for the case that the reporting is triggered by DCI, the funtionality activation delay is Tmeasurement period for inference.  Observation 1: RAN1 agreed to introduce new RRC parameter for CSI report configuration to distinguish CSI report of AI-CSI prediction and non-AI CSI prediction.  Proposal 4: for CSI prediction, it is proposed to define deactivation delay, considering that RAN1 agreed to introduce new RRC parameter for CSI report configuration to distinguish CSI report of AI-CSI prediction and non-AI CSI prediction.  Proposal 5: for AI/ML based beam management, AI/ML based performance requirement should not be worse than the legacy measurement requirements. |
| R4-2510335 | Vivo | Proposal 1: For performance monitoring reporting delay for BM-Case 1, the correasponding delay is defined as  Tmoitoring\_period\_BM = Nmonitor\* TL1-RSRP\_Prediction\_Period\_CSI-RS,  Where Nmonitor is the configured number of the latest transmission occasion(s) of monitoring resources with linked inference report no later than CSI reference resource corresponding to the CSI report for monitoring。  Proposal 2：For performance monitoring reporting delay for BM-Case 2, the specific delay definition can be deferred until RAN4 determines whether to specify RRM requirements for Case 2 (or which requirements to define) before further discussion.  Proposal 3: For Performance Monitoring Reporting Accuracy, it camn be defined in the test as: The rate of correct report quantity RS-PAI N\_p observed during repeated tests, where the report quantity RS-PAI follows RAN1's definition. Specifically, for BM-Case 2, the rate of correct report quantity RS-PAI reporting should be evaluated per time instance (if defined).  Proposal 4: RAN4 to discuss Type 3 performance monitoring for UE assisted performance monitoring of CSI prediction. RAN4 to define the delay requirement and accuracy requirement of SGCS A and SGCS B.  The delay for performance monitoring is only for one instance.  In RAN1 agreement, SGCS A is calculated based on predicted CSI for one inference reporting, and ground truth CSI, SGCS B is based on ground truth CSI and CSI (non-predicted) corresponding to the latest CSI-RS transmission occasion not later than CSI reference resource of the inference reporting instance. |
| R4-2510804 | Oppo | Proposal 1: Delay requirements for inference/monitoring reporting are necessary. Both measurement delay and AI/ML based processing delay before reporting should be included when defining reporting timelines. Detailed delay values for AI/ML processing remain FFS (depending on RAN1 agreements, e.g., on related UE capabilities).  Proposal 2: For AI/ML-based predictive use cases(time-domain CSI prediction, time-domain beam prediction), the reporting delay can be relaxed, i.e., when the prediction time span long, UE does not need to complete reporting immediately.  Proposal 3: RAN4 should define activation delay to prevent performance degradation caused by overly slow activation of AI/ML functionalities/models.  Proposal 4: RAN4 does not need to define delay requirements for deactivation.  Proposal 5: When specifying delay requirements for model switching, following two cases should be taken into account:  Case1: Under constrained computing resources, if insufficient resources are available to activate Model B immediately, the switching process may require first deactivating Model A to free up sufficient computing capacity before activating Model B, introducing additional timing constraints.  Case2: With sufficient computing resources, Model A deactivation and Model B activation can occur simultaneously, minimizing switching delay.  Proposal 6: Stability of performance monitoring and decision-making mechanism should be considered to mitigate the impact of random effects on monitoring outcomes. |
| R4-2510871 | Huawei, HiSilicon | Observation 1: Different delay requirements are identified for different performance monitoring reporting types.  Proposal 1: RAN4 defines different delay requirements for periodic, aperiodic and semi-persistent monitoring reporting.  Observation 2: NW controls when the uplink transmission resource for AI-based reporting is not available.  Proposal 2: RAN4 not to define deactivation requirements.  Proposal 3: If a mixed dataset is created for testing generalization, the mixed dataset should be static. |
| R4-2511150 | Ericsson | • 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 1 (post-deployment management based on LCM): Continue to work on monitoring to understand how useful and interoperable it can be.  • Proposal 2 (LCM): Further details on functionality activation delay requirement for AI/ML beam management can be discussed under the beam management use case.  • Proposal 3 (LCM): Functionality activation delay requirement is defined for the CSI prediction use case.  • Proposal 4 (LCM): To define functionality activation delay requirement for the CSI prediction use case, the AI/ML beam management framework can be reused. |
| R4-2511569 | Qualcomm | Observation 1: RAN4 has already agreed that activation delay starts at the reception of DCI or MAC-CE.  Observation 2: Activation delay can only be checked based on the timing of UE’s inference report.  Observation 3: The required measurement delay of a particular AI-ML model should be same during first and subsequent AP/SP reporting of predicted L1-RSRP.  Observation 4: UE may need to run many AI-ML models in time. UE cannot store all different AI-ML models in on-chip memory.  Observation 5: A typical UE would keep different AI-ML models in off-chip memory and load the model every time it runs inference and generate report.  • With above approach, inference timeline of a particular AI-ML model should be same during first and subsequent AP/SP reporting of predicted L1-RSRP, too.  Observation 6: RAN1 does not differentiate inference time between first and subsequent aperiodic reporting of predicted L1-RSRP.  Observation 7: It is not clear what UE is supposed to report after deactivating an AI-ML functionality.  Proposal 1: RAN4 agrees inference delay to be same between first and subsequent aperiodic reporting of predicted L1-RSRP after RRC reconfiguration  • Note: This means that the required time separation between following channels should be same for both first and subsequent aperiodic reporting of predicted L1-RSRP after RRC reconfiguration  o DCI and inference report  o SetB reference signal and inference report  Proposal 2: RAN4 deprioritizes defining deactivation delay requirement. |

## Open issues summary

**Issue 1-1: LCM for beam management**

Agreement from RAN4#114bis:

* RAN4 has the following understanding and need further check whether the understanding is aligned with RAN1 and RAN2:
  + - * **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

WF from RAN4#115:

Activation delay:

* Activation delay requirements created for the periodic, semi-persistent and aperiodic CSI reporting for beam prediction
  + It is FFS for CSI prediction
* For Option A, RAN4 will send LS to RAN2 to clarify the related mechanism.
* For semi-persistent CSI reporting
  + Activation delay starts at the reception of the MAC-CE/DCI
* For aperiodic CSI reporting
  + Activation delay starts at the reception of the DCI

Switching delay and Fallback delay:

* Deprioritize the discussion on switching and fallback delay requirements and focus on activation/deactivation delay requirement, including both AI/nonAI, discussion.

Monitoring reporting delay and accuracy:

Discuss how to define delay and accuracy requirements for UE sided monitoring where the UE reports the metric:

* BM-Case1 and possibly BM-Case 2

o Performance monitoring type 1 option 2

* If it proves to be not possible to define, then reverse the agreement
* Note: For BM-Case2, it may not be completed in Rel-19. If it is not covered in Rel-19, there would not be a monitoring delay requirement. If it would be in Rel-19, a monitoring delay requirement is needed.
* Proposals
  + Further discuss activation delay requirement under the beam management use case (Ericsson proposal 2)
  + Create delay requirements for LCM monitoring/reporting Oppo proposal 1)
  + Activation delay for periodic CSI reporting is defined as follows (Vivo proposal 1)
  + Upon reporting applicable functionalities via RRCReconfigurationComplete in step 4 (for Option A) or upon the reception of CSI-ReportConfig for inference configuration in RRCReconfiguration message of step 5 (for Option B) in slot n, the UE shall be capable to transmit valid CSI inference report and apply actions for the functionality being activated no later than in slot n+(T\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length) , where:
  + Tactivation\_time is the activation delay of the functionality in milliseconds, including preprocessing for inference (e.g., loading required information for inference).
  + TCSI\_reporting is the delay (in ms) including uncertainty in acquiring the first available downlink CSI reference resource, UE processing time for CSI reporting and uncertainty in acquiring the first available CSI reporting resources.
  + Note: The definition of the above delay is based on the working assumption that "activation begins only after the UE has reported applicability reporting and the network has sent the complete inference configuration."
  + activation delay is defined as below:
  + Upon reiceiving MAC-CE triggering activation command in slot n, the UE shall be capable to transmit valid CSI inference report and apply actions for the functionality being activated no later than in slot n+(〖T\_HARQ+[3]ms+T〗\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length) , where:
  + Tactivation\_time is the functionality activation delay in milliseconds, including preprocessing for inference (e.g., loading required information for inference).
  + TCSI\_reporting is the delay (in ms) including uncertainty in acquiring the first available downlink CSI reference resource, UE processing time for CSI reporting and uncertainty in acquiring the first available CSI reporting resources.
* Activation delay requirements should be defined to prevent degradation (Oppo proposal 3)
* For DCI based semi-persistent CSI reporting, the activation delay definition is similar to periodic (Vivo proposal 3)
* Activation delay is from RRC complete message until first valid report (Apple proposal 9, Nokia proposal 8-9)
  + Endpoint for activation is sending the inference report (CMCC proposal 1)
* Activation does not need to be finished before sending RRCReconfigurationComplete (Xiaomi proposal 1)
* Discuss option B first for periodic CSI reporting (Mediatek proposal 1)
* Activation delay starts from RRCReconfigurationComplete after step 5 for option B (Mediatek proposal 2)
* Divide activation time into time needed for activation and time needed for reporting (Mediatek proposal 3)
  + Ask RAN2 to create signalling for reporting of activation time for the UE (Mediatek proposal 4)
* Assume functionality partially prepared before MAC-CE triggering for semi-persitent and aperiodic reporting. A second activation time is the time needed from receiving MAC\_CE (Mediatek proposal 5-6)
  + The activation time after MAC\_CE is THARQ+3msec for MAC\_CE decoding plus up to 2msec depending on UE capability (MEdiatek proposal 7)
  + The second activation time is THARQ+3msec+measurement time for inference (CMCC proposal 2)
* For DCI activation, the activation time is the inference measurement time (CMCC proposal 3)
* Deactivation delay is from deactivation command to the end of reporting (Apple proposal 9)
* Do not introduce deactivation requirements (vivo proposal 5, Xiaomi proposal 2, Nokia proposal 10, Oppo proposal 4, Huawei proposal 2, Qualcomm proposal 2)
* Deactivation requirements only needed if e.g. the network needs to know when it can release resources for measurement. No interruptions expected during this time. Requirement may not be needed. (Mediatek proposal 8)
* Do not define performance monitoring delay for BM (Xiaomi proposal 3)
* Define different performance monitoring delay requirements for periodic, aperiodic and semi-persistant reporting (Huawei proposal 1)
* Define monitoring delay requirements (Nokia proposal 5)
  + SStarting from RRCReportComplete, ending when the first monitoring metric is sent
* Define monitoring delay requirements as (Vivo proposal 1):

Tmoitoring\_period\_BM = Nmonitor\* TL1-RSRP\_Prediction\_Period\_CSI-RS,

* Where Nmonitor is the configured number of the latest transmission occasion(s) of monitoring resources with linked inference report no later than CSI reference resource corresponding to the CSI report for monitoring
* Define a beam prediction accuracy (Xiaomi proposals 4-5, vivo proposal 3):
  + Beam prediction accuracy is defined as the ratio of the number of correctly predicted beam instances to the total number of evaluated monitoring occasions over a configured time window, i.e. Np/N.
  + Where,
  + - N is the number of latest CSI-RS or SSB transmission occasions used for performance evaluation, as configured by the network.
  + - Np is the number of those occasions for which the beam prediction is considered accurate according to the matching rule (e.g., at least one of the Top-K predicted beams matches one of the Top-M strongest beams measured based on L1-RSRP).The values of K, M, and N are network-configurable parameters.
* Define a requirement for monitoring report accuracy (Nokia proposal 4)
  + Test in at least static conditions
* For BM case two, discuss whether it is possible to cover all of the following aspects or whether it is not possible to complete requirements in Rel-19 (vivo proposal 6, observation 2)
* Aspect 1: Prediction delay requirement for BM-Case 2
* Aspect 2: TCI state-related (including TCI state known condition and MAC-based TCI state switch delay)
* Aspect 3: Metrics/KPIs for BM-Case 2 and corresponding performance evaluation
* Aspect 4: Performance monitoring-related (including delay requirement and accuracy requirement)
* For BM case two, decide whether to specify RRM requirements; do not start work on delay requirement until the decision is made (Vivo proposal 2)
* RAN4 assumes that inference delay is the same between the first and subsequent aperiodic reporting of L1-RSRP (Qualcomm proposal 1)
* Recommended WF
  + For activation delay, define activation delay as follows:
    - For option A
      * For periodic CSI reporting, the time from sending RRCReconfigurationComplete in step 4 to transmitting a valid CSI inference report.
      * For semi-persistent or aperiodic reporting, the time from receiving MAC-CE or DCI to transmitting a valid CSI inference report.
    - For option B
      * For periodic CSI reporting, the time from receiving RRCReconfiguration in step 5 to transmitting a valid CSI inference report.
      * For semi-persistent or aperiodic reporting, the time from receiving MAC-CE or DCI to transmitting a valid CSI inference report.
    - For both options
      * For periodic reporting delay is (T\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length)
      * For aperiodic reporting delay is (〖T\_HARQ+[3]ms+T〗\_(activation\_time)+T\_(CSI\_Reporting))/(NR slot length)
      * T\_activation time is defined below
      * T\_CSI\_reporting is the maximum time needed to acquire the first available CSI resource.
    - Discuss whether T\_activation time should be fixed or can be signalled from the UE.
  + Do not create deactivation delay requirements
  + Create performance monitoring delay requirements

Tmoitoring\_period\_BM = Nmonitor\* TL1-RSRP\_Prediction\_Period\_CSI-RS,

* Where Nmonitor is the configured number of the latest transmission occasion(s) of monitoring resources with linked inference report no later than CSI reference resource corresponding to the CSI report for monitoring
  + Define beam prediction accuracy requirements as follows:
    - the number of correctly predicted beam instances to the total number of evaluated monitoring occasions over a configured time window, i.e. Np/N.

Where,

* + - - N is the number of latest CSI-RS or SSB transmission occasions used for performance evaluation, as configured by the network.
    - - Np is the number of those occasions for which the beam prediction is considered accurate according to the matching rule (e.g., at least one of the Top-K predicted beams matches one of the Top-M strongest beams measured based on L1-RSRP).The values of K, M, and N are network-configurable parameters.
  + Discuss whether to create RRM requirements for beam management case 2

**Issue 1-2: LCM for CSI prediction**

Agreement from RAN4#114bis:

* RAN4 has the following understanding and need further check whether the understanding is aligned with RAN1 and RAN2:
  + - * **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

WF from RAN4#115:

CSI prediction:

* CSI prediction, the details are subject to RAN1 decision
* RAN4 will further study the feasibility to specify the corresponding delay and/or accuracy requirements for UE sided monitoring

**Way forward (not agreements)**

Deactivation delay for CSI prediction:

* The following questions were noted for further discussion
  + Can NW differentiate between the predicted CSI reporting and the measured CSI reporting?
  + In case of switch between AI, is there existing reporting to differentiate the reports before and after the deactivation.
* Proposals
  + Introduce activation delay requirements for periodic, semi-persistent and aperiodic CSI reporting (vivo proposal 4, Oppo proposal 1)
  + Introduce activation delay requirements using the same framework as beam management (Ericsson proposals 3-4)
  + Do not introduce deactivation requirements (vivo proposal 5, Nokia proposal 10, Oppo proposal 4, Huawei proposal 2, Qualcomm proposal 2)
  + Define deactivation requirements (CMCC proposal 4)
  + Define reporting accuracy requirements (Nokia proposal 6)
    - For SGCS1 values
  + Define monitoring reporting delay requirements (Nokia proposal 7)
    - From RRCReconfigurationComplete to first monitoring report
  + For CSI prediction, the reporting delay can be long (Oppo proposal 2)
* Recommended WF
  + Define activation delay requirements
    - Use the same framework as for beam management (if agreed above).
      * Check whether the activation time should be assumed the same or not.
  + Do not define deactivation delay requirements
  + Define monitoring delay requirements
    - Use the same framework as for beam management (if agreed above)
      * Check whether the monitoring delay can be different to beam management
  + Define requirements for monitoring report accuracy

**Issue 1-3: LCM general**

* Proposals
  + Derive a test to ensure LCM decision accuracy; for example, metrics improve (NTU proposal 1)
  + Create delay requirements for LCM monitoring/reporting Oppo proposal 1)
  + Create different delay classes for activation and feedback delay depending on how time critical the AI functionalities are (Apple proposal 5)
  + General principles for activation/deactivation/fallback (Apple proposal 6):
    - For RRC based, delay is from activation of the RRC command to RRC complete message
    - For MAC CE/DCI based starting at the reception of control signalling. A signal from the UE is needed (such as acknowledgement or CSI feedback) to define a completion time.
    - The end time can be the first valid measurement (Apple proposal 8)
  + The UE shall be able to signal if KPIs for runtime operation fall outside of a certain window (Apple proposal 7)
  + Activation delay requirements should be defined to prevent degradation (Oppo proposal 3)
  + No need to create de-activation delay requirements (Oppo proposal 4, Huawei proposal 2, Qualcomm proposal 2)
  + When considering switching requirements, consider two scenarios; models need to switch off/on sequentially and models can be switched off/on in parallel (Oppo proposal 5)
* Recommended WF:
  + For LCM decision accuracy, it seems to be out of the scope of RAN4 as it would be a test on the management entity performing the LCM.
    - Clarification of any misunderstanding welcome
  + Check whether it is possible to reach a general decision (by default applicable to all use cases) on the following:
    - Activation delay requirements will be created
    - Monitoring delay/fallback delay requirements will be created
    - Discuss whether there can be different classes of activation/fallback delay depending on the AI functionality.
    - The following general principles can be applied to AI functionality that impacts CSI reporting:
      * For RRC based, delay is from activation of the RRC command to RRC complete message
      * For MAC CE/DCI based starting at the reception of control signalling. A signal from the UE is needed (such as acknowledgement or CSI feedback) to define a completion time.
      * The end time can be the first valid measurement (Apple proposal 8)
    - No need to create deactivation requirements for AI functionality.
    - Discuss whether to further consider switching requirements

Note: Before discussing these general principles for LCM, the LCM for beam management and CSI will be discussed first. If agreements are reached for CSI and BM, this discussion will be about whether the assumptions can be taken as a more general baseline for AI LCM.

**Issue 1-4: LCM framework for 2-sided model**

* Proposals
  + Use the single sided UE LCM framework as a starting point and then study if it can be re-applied for 2-sided (Apple proposal 2, Nokia proposal 1)
* Recommended WF
  + Agree that for 2-sided, the Rel-20 WI could consider the LCM for UE one sided as a starting point

**Issue 1-5: Post-deployment testing / model update**

Agreement from RAN4#114bis:

* Rename “option 1” as “Post-deployment pre-activation functionality/configuration update testing”
* Rename “option 2” as “Post-deployment post-activation functionality testing based on performance monitoring”

WF from RAN4#115:

Option 2 (Post-deployment post-activation functionality testing based on performance monitoring) is prioritized.

* it should be further investigated/established that reliable and testable monitoring procedures can be established for all use cases
* Proposals
  + Develop a hybrid approach, for which major updates undergo testing in a lab or on a device, whilst minor updates are subject to monitoring (Apple proposal 1)
  + Devise criteria for minor and major updates; e.g. scope of model training, architecture change, new channel profile or dataset etc. (Apple proposal 3)
  + Monitoring for updates according to option 2 shall be based on a “shadow” model that creates inference reports, but is not being actively used (Apple proposal 2, Nokia proposal 1)
  + No new requirements needed for “test active” functionality; it is monitoring (Nokia proposal 2)
  + Consider how to optimize post deployment monitoring across multiple UEs (Nokia proposal 3)
  + KPI envelopes for monitoring are related to the scenario and e.g. previous model behaviour or testing metrics (Apple Proposal 4)
  + There is a need to know when an updated behaviour is effective from. Discuss an uplink “effective from” signalling, with rules on delivery and classification (KTL observations, proposal 1, 2)
  + Differentiate between monitoring for post-deployment update handling and operational handling (KTL observations and proposal 3)
  + Consider an indicator of when monitoring is for validation or for runtime (KTL proposal 4)
* Recommended WF
  + For option 2, discuss whether there is a need to create a “shadow” or “test active” possibility to run a model whose output is not actively being used and produce monitoring reports
  + Discuss whether to introduce a signalling of when behaviour is effective from
  + Discuss whether to differentiate the monitoring used for post-deployment handling from that used for operational handling
  + Considering option 2 and option 1 (note that option 2 is prioritized following agreements at RAN4#115):
    - Discuss whether to differentiate “major” and “minor” updates and apply pre-testing for major updates, monitoring for minor updates.
      * For major updates, discuss whether testing is on device or in a lab.

**Issue 1-6: Post-deployment testing / model update for 2-sided**

* Proposals
  + Clarify whether the agreement on post-deployment testing to prioritize option 2 (monitoring) also applies for 2-sided (Nokia proposal 6)
* Recommended WF
  + Agree to agree the previous agreement also for 2-sided.

**Issue 1-7: Performance monitoring requirements for 2-sided**

* Proposals
  + For network sided monitoring, use legacy delay requirements (no new requirements) (Apple proposal 3, Nokia proposal 2)
  + For UE sided monitoring, derive new delay requirements (Apple proposal 3)
  + For UE sided monitoring, discuss whether to derive delay and also accuracy requirements (Nokia proposal 3)
* Recommended WF
  + Check if possible to agree no new delay requirements for NW sided monitoring.
  + Check if possible to agree in principle to create delay requirements for UE sided reporting.

**Issue 1-8: Generalization and scalability**

WF from RAN4#114bis:

**Agreements**

* 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.
* 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.

WF from RAN4#115:

The high level and general guidelines for the generalization test are provided as follow. The exact decisions are to be made case by case.

* Study and , if feasible, define requirements for each AI/ML functionality
* Define one test per UE capability as a minimum
  + FFS >1 test per UE capability
* Define a minimum set of test configurations (including NW sided conditions if any) as mandatory for testing of AI/ML-enabled Feature

Potential areas to consider for generalization testing:

* gNB array parameters
  + port layouts, array size, antenna virtualization
* propagation conditions
* Deployment scenarios
  + Carrier frequencies
  + Speeds
  + Indoor/outdoor
  + Bandwidths
* SNR
* Proposals
  + Consider the following for generalization testing (Nokia proposal 4)
    - Deployment scenarios
    - Indoor/outdoor UE distributions
    - Carrier frequencies
    - TxRU mappings
  + Discuss the need for scalability testing (Nokia proposal 5)
  + Check the granularity of the UE feature list. Consider >1 test per UE capability if there are several components per UE feature (CMCC proposals 1-2)
  + Study generalization at the level of AI/ML feature groups (Nokia proposal 11)
  + Do not consider mixed model dataset for generalization testing (Nokia proposal 12)
  + If a mixed dataset is considered it should be static (Huawei proposal 3)
* Recommended WF
  + For 2-sided generalization testing:
    - Deployment scenarios
    - Indoor/outdoor UE distributions
    - Carrier frequencies
    - TxRU mappings
  + For 2-sided generalization testing, consider testing scalability
  + For generalization testing in general, there may be a need to consider different components of capability. It is preferable to discuss this on a use-case basis once information on capability granularity is agreed in other WG.
  + For the mixed dataset, address the question of what the purpose of the generalization testing is; is it exposure to a variety of scenarios or also changes in scenarios.

**Issue 1-9: Principle for setting requirements to ensure superior performance of AI**

* Proposals
  + Consider a composite metric considering performance gain and/or system efficiency improvement (Apple proposal 1)
  + For CSI prediction, discuss more specifically the following (Huawei proposal 1)
    - 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.
  + For beam management, use L1-RSRP requirement for NW sided model (Huawei proposal 2)
  + FFS for UE sided prediction accuracy for beam management (Huawei proposal 3)
  + For positioning cases 3a, 3b, re-use legacy requirement for timing information, paired timing information and power information (Huawei proposal 4)
  + For BM RSRP prediction accuracy, measurement error may influence whether AI requirements can exceed “legacy” (Xiaomi proposal 6)
  + For BM, AI/ML performance should not be worse than legacy (CMCC proposal 5)
* Recommended WF
  + Discuss whether to consider the proposal for a composite metric of system overhead and performance gain, and how this might map to actual RAN4 requirements.
    - Check if a general conclusion is possible or whether it should be considered in individual use-cases.
  + For CSI prediction, there is a need to discuss the assumed CSI feedback and the requirement threshold as part of CSI prediction performance.
    - Check if possible to agree whether the aim is to increase throughput or reduce CSI overhead in CSI performance discussion.
  + For beam management, confirm in the beam management discussion whether for WN sided model, the existing L1-RSRP accuracy can be used for the UE reporting.
  + For positioning, check in the positioning thread whether for 3a/3b, the legacy requirement can be re-used for timing information, paired timing information and power information.

# Topic #2: Two-sided CSI compression

This topic handles the 2-sided CSI study item.

The discussion should try to reach conclusions on the remaining simulation on alternative backbone and low complexity encoders. Some remaining issues for the SI include observations on field data, encoder/decoder selection criteria and a comparison of option 3 and option 4 (although it should be noted that option 3 is already agreed for the Rel-20 WI). Some proposals and discussion are also provided on next steps for the Rel-20 WI.

During the meeting, the TP to the TR and the simulation summary should be discussed, mainly offline, in order to prepare them for approval.

## Companies’ contributions summary

|  |  |  |
| --- | --- | --- |
| **T-doc number** | **Company** | **Proposals / Observations** |
| R4-2509241 | NTU | Proposal 1: Consider the following as (test/reference) encoder/decoder selection criterion:  Achievable performance metrics: SGCS or NMSE  Complexity: including flops or model storage size, can consider to set an upper bound  Robustness (based on performance metrics): whether the decoder can achieve satisfactory performance when connecting to encoder with different structures with the decoder, but trained with the decoder’s input and output dataset  Proposal 2: For all the candidate decoders, evaluate the robustness metrics based on test repeatability and interoperability by the following procedure  Companies contribute encoders trained with the candidate decoders (one encoder per candidate decoder)  Connect a encoder to the corresponding decoder, run the test based on the agree test dataset and record the SGCS  Repeat step 2 for all the encoder and decoder pairs, record all the SGCS  For each candidate decoder, capture the variation of the SGCSs when connecting to different encoders, and evaluate the variation together with performance metrics and complexity  Proposal 3: 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 δ via simulation study (setup following option 3 discussion), and check whether it’s feasible to derive a test decoder satisfying  ∑\_(z∈Z)▒f(g\_ref (z),g\_test (z)) <δ,  where Z is the set of decoder inputs (latent messages) in the specified dataset (from step 2), and f is the chosen loss/similarity function g\_ref (z) is the test decoder output and g\_test (z) is from the dataset.  4a-2: RAN4 studies whether it is feasible to derive a test decoder satisfying  ‖g\_ref (z)-g\_test (z)‖=(min)┬(y∈Y\_test )⁡‖g\_ref (z)-y‖ for all z∈Z  where the notations are the same as 4a-1, and Y\_test is the test decoder output space.  (Option 4b: specify reference encoder)  4b-1: RAN4 derive the threshold δ 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  ∑\_(x∈X)▒f(g\_ref (z(x)),g\_test (z(x))) <δ,  where z(x) is the reference encoder output (latent message) of encoder input x, and X 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  ‖g\_ref (z)-g\_test (z)‖=(min)┬(y∈Y\_test )⁡‖g\_ref (z)-y ̂ ‖ for all z∈Z  where Y\_test 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.  Observation 1: Option 4a-1 and 4b-1 require both TE vendors and UE vendors to train the test decoder and DUT encoder with the dataset or reference encoder, while option 4a-2 and 4b-2 allow TE vendors and UE vendors to train the test decoder and DUT encoder with arbitrary dataset or encoders/decoders as long as the specified ordering requirements are satisfied. |
| R4-2509301 | CATT | Observation 1: In 1-to1 joint training and separate encoder training with fixed decoder, low complexity encoders with CNN/transformer backbone have comparable performance to that of Encoder 1. MLP encoder has the lowest performance and a 5% SGCS degradation compared to others.  Observation 2: In N-to-1 joint training, low complexity encoders with CNN/transformer/MLP backbone have comparable performance to that of Encoder 1, where the SGCS spans around 2.5%.  Observation 3: N-to-1 joint training provides performance generalization and robustness when different backbones/structures are used for encoders, compared with 1-to-1 training. |
| R4-2509430 | Apple | Observation 1 Option 4’s core benefit is flexibility without abandoning common design rules. Instead of specifying a single frozen decoder (Option 3), each TE vendor can train its own test decoder optimizing latency/memory footprint while still following 3GPP defined anchors. A second gain is future proofing and maintainability. Because the TE decoder is retrainable, refreshing to new channel models, can be done by distributing new data or an updated reference encoder no need to reopen the spec to publish a brand new frozen decoder binary. Third, Option 4 empowers vendor side optimization and IP protection. UE vendors can use an “own” decoder during encoder training to better match their proprietary front end processing.  Observation 2 Option 4b gives a fixed latent space (via the frozen reference encoder), while 4a only gives a set of labels that depends on how those labels were produced. With 4b, every TE vendor trains its decoder to invert the same encoder output distribution, so partial-structure freedom doesn’t break interoperability. In 4a, unless the labelled dataset was generated by that exact encoder, different “label sources” or updates can shift the latent space and re open the mismatch problem. If those labels weren’t produced by the frozen reference encoder, then every TE vendor will train its decoder toward a different latent space.  Observation 3 Option 4a collapses to Option 4b when the dataset is nothing more than the reference encoder’s outputs. Concretely, if every label c in the 4a set is generated as c=f\_ref(H) by a single, fixed reference encoder (same weights, pre/post processing, quantization) and TE vendors are required to train on that dataset, then the latent space is anchored by f\_ref just as in 4b. At that point: The encoder is effectively standardized, only hidden inside the dataset. TE decoders are all learning to invert the same code distribution. Thus the “dataset-anchored” option (4a) becomes functionally identical to the “encoder-anchored” option (4b)  Observation 4 Option 3 and Option 4 become effectively identical when all the flexibility intended for Option 4 is removed. This happens if: (1) the TE test decoder is fully specified and identical for all vendors, (2) everyone uses the same mixed/reference dataset, (3) no vendor-specific “own” decoder or retraining is used at TE , and (4) any small freedoms in training setup (e.g., optimizer, seeds, pruning) are fixed or have no impact on the final decoder. In that case, Option 4’s “design-your-own decoder” path reduces to Option 3’s “use this frozen decoder,” with the same latent space, same weights, and same behavior. The two also converge if Option 4b’s frozen reference encoder and dataset yield a single common decoder that vendors simply adopt. Similarly, if Option 4a’s labelled dataset is generated by the same reference encoder and the decoder’s backbone and architecture are frozen, there’s no remaining design freedom, making 4a essentially the same as Option 3  Observation 5 Option 4b lets refresh the training components for test decoder (dataset and/or frozen reference encoder) instead of re-freezing and re-specifying a new decoder every time channels change as in Option 3. TE vendors can retrain their test decoders on the updated data within the same structural envelope, so the spec stays stable while the models adapt. Option 3 would require publishing a brand new frozen decoder (and re-certifying everyone) whenever the channel model set evolves  Based on the discussion in the previous sections we propose the following:  Proposal 1 RAN4 should standardize a single frozen backbone test decoder and one associated reference encoder, trained on a sufficiently diverse dataset covering agreed frequency bands, environments, mobility profiles, antenna configurations, and SNR ranges. Additional full decoders shall only be added if objective evidence (e.g., agreed NMSE or SGCS spread thresholds across distinct requirement scenarios) proves that a single backbone cannot meet performance requirements. This ensures interoperability while avoiding unnecessary proliferation of large decoder models.  Proposal 2 Adopt Backbone + Scenario-Specific Adapter Framework  For scenarios where performance with the single backbone falls short, RAN4 should adopt a backbone + adapter architecture. The frozen backbone defines a common latent space for interoperability, while small, scenario-specific adapters are trained using targeted datasets to specialize the backbone. The backbone remains unchanged across scenarios, ensuring stability; only adapters are trained and validated, minimizing storage, maintenance.  Proposal 3 On-Demand Adapter Generation Using Standardized Datasets  Instead of storing multiple pre-trained adapters, RAN4 should specify that new adapters be generated on demand during certification. For each new scenario, only the scenario-specific dataset and simulation assumptions are standardized. The adapter is trained on this dataset at certification time and combined with the frozen backbone to form the test decoder for that scenario. This approach eliminates per-scenario versioning overhead, reduces TE storage requirements, and ensures scalability when supporting future channel models or deployment conditions.  Proposal 4  To determine if more than one test decoder is needed, we propose a simulation framework for Test-Decoder Deployment Variability Assessment. Three model families are trained: (i) a universal model, (ii) standalone per-scenario models, and (iii) a hybrid with a shared backbone plus scenario-specific experts. Scenarios vary in carrier frequency, antenna configuration, user speed, environment, and SNR; evaluation is on both seen and unseen scenarios. Comparing “diagonal” errors (model tested on its own scenario) with “off-diagonal” errors (tested on other scenarios) reveals when multiple decoders are necessary and whether compact expert heads can match standalone model accuracy with better scalability.  Proposal 5 RAN4 shall retain a single reference encoder per frozen test decoder, selecting the lowest complexity encoder that still meets the agreed performance target; additional reference encoders would be justified only if RAN4 explicitly defines distinct performance tiers tied to UE capability classes.  Proposal 6 If RAN4 defines distinct performance tiers, RAN4 needs to set the exact convergence metric (SGCS, NMSE) define the capability tiers and their minimum performance floors  Proposal 7 Keep Option 3 as the baseline while maturing Option 4b, with pre agreed exit criteria to drop one path later. RAN4 should standardize a single frozen test decoder (Option 3) and its one reference encoder, guaranteeing immediate, low risk interoperability. In parallel, RAN4 should formalize Option 4b TE designed test decoders trained against a frozen reference encoder (and common/aligned dataset) under a partially specified structure to gain flexibility for future channel models, hardware targets, and innovation. Option 4a is retained only insofar as its labelled dataset is produced by that same frozen encoder; otherwise it should be folded into 4b.  Proposal 8 To prevent indefinite dual maintenance, RAN4 shall define quantitative “convergence” triggers before launching the joint trial: for example, SGCS (or NMSE) spread across vendors and scenarios ≤ 5 % (or ≤ 1 dB) and no loss of interoperability in cross-vendor tests compared to the Option 3 baseline. If Option 4b meets these thresholds, announce a sunset plan for Option 3 at a specific release boundary. If it does not, Option 3 remains the normative path and Option 4b reverts to an informative/experimental track until the gaps are closed.  Proposal 9 For training the test decoder a standardized dataset collection ensures the test decoder is trained on data that realistically reflects wireless channel scenarios, producing a representative and interoperable latent space across vendors. RAN4 shall define a simulation framework covering frequency bands, environments, mobility, antenna setups, and SNR distributions. All companies must follow these simulation agreements, with dataset alignment verified via metrics like power spectral entropy (PSE). Proprietary datasets must demonstrate alignment or be adjusted accordingly |
| R4-2509656 | Mediatek | Observation #1: Option 3 Track 2 SGCS testing 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 (joint training).  Observation #2: Testing results with frozen pretrained decoder (trained with mixed dataset) using lower complexity CNN Encoder are slightly worse compared to high complexity Encoder when trained with Mixed dataset.  Observation #3: Testing results with frozen pretrained decoder (trained with mixed dataset) using lower complexity Transformer Encoder are slightly worse compared to high complexity Encoder when trained with Mixed dataset.  Observation #4: Testing results with frozen pretrained decoder (trained with mixed dataset) using lower complexity MLP Encoder are clearly worse compared to high complexity Encoder when trained with Mixed dataset.  Observation #5: Testing results with frozen pretrained decoder (trained with mixed dataset) using lower complexity Transformer and MLP Encoders are clearly worse compared to high complexity Encoder when trained with MediaTek dataset and tested with Mixed dataset.  Observation #6: 7 decoders were successfully used for evaluation.  Observation #7: Option 4 SGCS results are comparable to SGCS results in Option 3 Track 2.  Observation #8: Performance difference between high and low complexity CNN Encoders depends on used training and testing datasets (varies between 0.01 to 0.05).  Observation #9: Performance difference between high complexity CNN Encoder and low complexity Transformer Encoder depends on used training and testing datasets (varies between 0.01 to 0.07).  Observation #10: Performance difference between high complexity CNN Encoder and low complexity MLP Encoder depends on used training and testing datasets (varies between 0.06 to 0.25). |
| R4-2509931 | Nokia | Observation 1: The separate training of the lower-complexity encoders with the high-complexity CNN-based frozen decoder may lead to a reduced SGCS performance compared to the separate training of the encoder and frozen decoder with the same backbone and complexity (i.e., high-complexity CNN-based).  Observation 2: The joint and separate training of the lower-complexity MLP-based encoder with the high-complexity CNN-based decoders give the lowest SGCS performance compared to the other joint and separate training results with the encoder backbones and complexities assumed in [R4-2508084].  Observation 3: Using multiple encoders with different backbones and complexities to derive the test decoder did not enhance the SGCS performance compared to using a single high-complexity CNN-based frozen decoder for the separate training.  Observation 4: A comparison between the SGCS performance of jointly trained low-complexity encoders with their matching decoders and that of a high-complexity decoder paired with an architecturally mismatched encoder confirms that decoder complexity is a key factor. A higher-complexity decoder consistently delivers superior SGCS performance, regardless of the encoder’s architectural backbone.  Proposal 1: RAN4 to select a high-complexity decoder model for the fully standardized test decoder.  Proposal 2: To define performance requirements and to verify test decoder, RAN4 needs to specify a reference encoder. It should be captured in RAN4 specifications.  Proposal 3: The reference encoder can be captured in RAN4 specifications using a link to a specified location of the model.  Proposal 4: RAN4 to capture one reference encoder in terms of complexity in RAN4 specifications.  Proposal 5: RAN4 to adopt the raw channel data format for the training dataset to capture in the three test decoder options.  Proposal 6: RAN4 to consider the following feasibility criteria for concluding on the feasibility of the test decoder options: An option can be feasible if it: (1) provides sufficient performance (e.g., ±5% from the SGCS obtained with Rel-16 eType2 codebook), and (2) can be standardized (e.g., based on the agreements about what to capture/specify for each option). |
| R4-2510166 | CMCC | Simulation results provided  The show SGCS drops with the N to 1 joint training, but is regained if new encoders are trained.  Around 0.04 SGCS drop for lower complexity encoders. |
| R4-2510184 | Korea Test Laboratory | Observation 1: In SGCS-4 Step-1, the choice of the anchor (Encoder-1) materially affects the SGCS of the other encoders.  Observation 2: With a limited budget (10 epochs in Step‑1 and 10 in Step‑2), the observed ranking is E4 (MLP) > E2 (Transformer) > E3 (CNN), but longer training may reverse it. To avoid over‑crediting lightweight models, report the learning-curves stability for each encoder(SGCS versus epochs), not just endpoints.  Observation 3: Relative to each vendor’s two‑side baseline, matched encoder–decoder pairs tend to improve while mismatched pairs tend to degrade. However, absolute rankings within a given decoder do not necessarily favor the matched vendor. Therefore, baseline-normalized reporting should be used to ensure fair cross-vendor comparison.  Observation 4: In line with WF [R4-2508051], convergence across encoders is typically within a few percent under aligned training/testing data, whereas misalignment can cause swings on the order of tens of percent  Proposal 1: Considering to require reporting of learning-curve stability (SGCS vs. epochs) for each encoder to rule out budget-induced ranking artifacts, since SGCS depends on the training-epoch budget.  Proposal 2: Discuss to adopt an anchor-fairness check as part of our evaluation framework. |
| R4-2510342 | Vivo | Observation 2: Encoder 2 (Transformer) can achieve best performance, while with low complexity.  Observation 3: SGCS-4 (2-step joint training) shows that one decoder can work with different encoder structures, when the training of decoder considers these different encoder structures.  Proposal 1: Reference encoder is to be specified in TS for Option 3/4a from Rel-19 SI perspective.  Proposal 2: Reference encoder will be developed and specified by RAN4.  Observation 4: Model specification for R20 WID includes:   Fully defined/specified reference encoder for testing decoder, for both 3a-1 and direction C.   Reference model is not the actual implemented model, but could be used for NW/UE to develop real decoder/encoder.   Fully specified test decoder.  Proposal 3: RAN4 job for check-point of inter-vendor training collaboration Direction C and Direction A, sub-option 3a-1:   Agree on steps to align the model structure and scalability, based on R19 RAN1/4 agreements, to show that RAN4 can fulfill the requirements from WID on reference model/structure.  Proposal 4: At least Transformer is considered as the backbone for reference encoder for 3a-1 and direction C. More backbones can be considered.  Proposal 5: For model structure scalability,   RAN1 confirms the scalability of model structure and parameters for direction A 3a-1 and direction C.   RAN4 to consider RAN1 scalability observations, and these simulations would not need to be repeated in RAN4.   RAN4 to pick the most straight forward and future-proof one among the alternatives.  Proposal 6: Similar methodology of RAN4 R19 study for test model options, can be carried on to Rel-20 WI to align on the reference model structure and scalability.   Step 0: Evaluation assumptions and mixed dataset based on Rel-19 can be reused.   Step 1: Align on the model backbone, model hyperparameters and the scalability solution.   Step 1-1: Proposed model structures (model backbone and model hyperparameters) are brought by companies, as well as proposed scalability solution.   Step 1-2: Check on performance alignment -> see simulation results from contributing companies.   Step 1-3: Share models (encoder or decoder or both)/datasets (training/testing/inference), if needed.   Step 1-4: Select one or more model structures.   Step 1-5: Select one or more scalability solutions. This step can be done in parallel with Step 1-4.   Step 2: Specify the model for the following work in WID. Where and how to specify may need further discussion.   Fully defined/specified reference encoder for testing decoder, for both 3a-1 and direction C, and   Fully specified test decoder.  Proposal 7: Using the mixed dataset for reference model training, including the mixing of SLS and LLS, while using the LLS dataset for RAN4 tests. Other mixing rules are not precluded.  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.  Observation 7: In field performance, the reference encoder performs better than reference decoder. |
| R4-2510810 | Oppo | Proposal 1: Principles to define test decoder(s)  • To meet the minimum performance requirement in RAN4 tests  • To be a simple design  • Consider different UE capabilities, including supported models, architectures, and computational complexities  Observation 1: While both RAN4 option 3 and RAN1 direction C may involve defining a test model or reference model, they address different issues and require separate solutions. The following key differences should be taken into account in RAN4 following works  • Motivation  • Model structure  • Testing data & Performance requirement  • Assumptions on generalization & Scalability consideration  Proposal 2: Regarding the inter-vendor training collaboration for two-sided AI/ML models, to fully defined/specified reference model (“Direction C”) with RAN1 scalability study outcome taken into account, how to take RAN1 requirements into account should clarified first. Following way forward could be discussed in RAN4 to check companies’ view.  • Option 1: RAN4 develops a standardized model based on RAN4’s testing requirements first, then RAN1 evaluates whether it can also be leveraged for RAN1 purposes.  • Option 2: RAN4 develops a standardized model that fulfills both RAN4’s testing requirements and RAN1’s commercial deployment requirements.  • Option 3: RAN4 develops two separate standardized models, one to meet RAN4’s testing needs and another one to fulfill RAN1’s commercial deployment requirements.  • For option2 and option3, RAN1’s requirements should be clarified and let RAN4 know.  Observation 2: The updated simulation results for low-complexity encoders with different backbones are shown in Table 3.  Observation 3: For Encoder-2 and Encoder-3, the obtained SGCS-2 is similar to SGCS-1, that shows a common decoder that is compatible to encoders with different backbones(at least for Transformer/CNN based lower-complexity encoder) and with different complexities. For Encoder-3, SGCS-2 is much lower than SGCS-1, but SGCS-4 is similar to SGCS-1.  Proposal 3: Low-complexity encoders with different backbones should be used for deriving the test decoder. |
| R4-2510837 | Ericsson | Observations:  Observation 1: For the agreed decoder architecture and training hyper parameters, SGCS-4 step 1 shows better SGCS compared with SGCS-0/3. That is, N-to-1 joint training gives a better performance than 1-to-1 joint training.  Observation 2: For the agreed decoder architecture and training hyper parameters, SGCS-4 step 2 shows better SGCS compared with SGCS-1/2. That is, encoders trained with a frozen decoder obtained from N-to-1 joint training have a better performance than the encoders trained with a frozen decoder obtained from 1-to-1 joint training of Encoder-1.  Observation 3: For the (low-complexity) decoder with convolution size of 32 and low complexity CNN encoders, SGCS-4 step 1 shows worse SGCS compared with SGCS-0/3. That is, N-to-1 joint training has a lower performance than 1-to-1 joint training.  Observation 4: For the (low-complexity) decoder with convolution size of 32, SGCS-4 step 2 for Encoder-5 only shows better SGCS compared with SGCS-2.  Observation 5: It is difficult to conclude the benefit of training the decoder with multiple encoder configurations and/or backbone.  Proposals:  Proposal 1: RAN4 needs to discuss how to capture the reference encoder and reference decoder structures (i.e., backbone, numbers of layers, type of layers and all description of the model) in Rel-20 WI.  Proposal 2: RAN4 needs to discuss how to capture the reference encoder and reference decoder parameters (i.e., model weights) in Rel-20 WI.  Proposal 3: RAN4 needs to revisit the dataset format including latent space in Rel-20 WI.  Proposal 4: RAN4 needs to discuss the test metric of PMI reporting requirements with two-sided CSI compression model in Rel-20 WI.  Proposal 5: RAN4 needs to discuss new encoder/decoder architecture in Rel-20 WI.  Proposal 6: RAN4 needs to introduce new dataset(s) ensuring the model generality in Rel-20 WI. |
| R4-2510878 | Huawei, HiSilicon | **Observation 1**: Comparing SGCS-1 (separate training for assumed CNN encoder with fixed CNN decoder) and SGCS-2 (separate training for low-complexity encoders with fixed CNN decoder), SGCS performance loss is observed.  **Observation 2**: Comparing SGCS-3 (joint training for low-complexity encoders with assumed CNN decoder) and SGCS-2 (separate training for low-complexity encoders with fixed CNN decoder), SGCS performance improvement is observed.  **Observation 3**: Comparing SGCS-3 (joint training for low-complexity encoders with assumed CNN decoder) and SGCS-0 (joint training for assumed CNN encoder with assumed CNN decoder), marginal SGCS performance loss is observed.  **Proposal 2**: Use low-complexity encoders to derive the common test decoder during WI.  **Observation 4**: Scalability aspects include number of Tx ports, bandwidth and subband size, CSI feedback payload size.  **Observation 5**: For each scalability aspect, several alternative solutions have been identified in RAN1.  **Observation 6**: The workload is extremely high if all combinations of alternatives for different scalability aspects are considered.  **Observation 7**: SGCS performance loss is around 0.023 on average for scalability over model input size with the baseline method for achieving model scalability proposed in Table 2.  **Observation 8**: SGCS performance loss is around 0.018 on average for scalability over model output size with the baseline method for achieving model scalability proposed in Table 2.  **Proposal 3**: For each scalability aspect, one alternative among all the identified alternatives is selected to form a baseline solution for deriving the scalable model.  **Proposal 4**: According to RAN1 simulation results on scalability study, the baseline solution is selected as the combination of the following alternatives in Table 2.  **Table 2**. Baseline solution for achieving model scalability   |  |  | | --- | --- | |  | Scalability solution | | Choice of token/feature dimension (if Transformer is agreed) | * Alt 1: Use subband as the token dimension and Tx port as a feature dimension   + The number of tokens varies with the number of subbands. | | Scalability over the feature dimension | * Alt2: a common embedding layer with padding (e.g., zero-padding or other techniques for padding values) | | Scalability over the token dimension | * Alt2: Padding at the input | | Scalability over payload configurations | * Alt2: truncation/masking of the output linear layer output |   **Observation 9**: Principles of designing the procedure to specify a scalable model during WI are summarized as follows.   * Mixed dataset generation: Align on different configurations for Tx port number and subband size. Companies will upload dataset with description of configurations. Align on configurations for CSI feedback payload size. * Model structures of both encoder and decoder: Select model structures based on companies’ contributions with an aligned scalability method. A reasonable complexity level range is defined for companies to propose their preferred models. Down-selection is performed if several models are very similar to each other. * Training procedure: If multiple encoders are considered for deriving the common test decoder, training procedure needs to be aligned, such as N-to-1 joint training and N-to-1 separate training. * Performance alignment metric: SGCS and throughput gain compared to non-AI under testing channel conditions are used. If models are aligned well but cannot achieve acceptable/testable performance gain, model structures need to be reselected. * Performance requirement definition: If multiple encoders are considered for deriving the common test decoder, select one encoder to be specified based on the criteria of the trade-off between throughput performance gain and complexity level.   **Proposal 5**: Steps for aligning the specified model during WI are summarized as follows.   * **Step 1**: Align on configurations for mixed dataset construction, taking scalability over Tx port number and subband size into account. * **Step 2**: Align on the model backbones and model structures based on companies’ contributions with an aligned scalability method, where model complexity is limited to a predefined range. One model structure is selected for decoder, while one or more model structures are selected for encoder.   + The baseline scalability method is as follows.  |  |  | | --- | --- | | Choice of token/feature dimension (if Transformer is agreed) | Use subband as the token dimension and Tx port as a feature dimension   * + The number of tokens varies with the number of subbands. | | Scalability over the feature dimension | A common embedding layer with padding (e.g., zero-padding or other techniques for padding values) | | Scalability over the token dimension | Padding at the input | | Scalability over payload configurations | Truncation/masking of the output linear layer output |  * **Step 3**: Align on training procedures (e.g., joint training and separate training, N-to-1 training if several encoders are used), hyperparameters and loss function definition. * **Step 4**: Share encoders and decoders and cross-check. * **Step 5**: Select one decoder with medium performance, that can achieve a testable performance gain under link-level testing channel conditions compared to non-AI.   + If no decoders meet the requirement, go back to Step 2 to reselect the model structures. * **Step 6**: Retrain encoders with the selected and fixed decoder. Share the encoder and cross-check.   + If performance is not aligned, go back to Step 5 to reselect the fixed decoder. * **Step 7**: Select one encoder from the retrained encoders which can achieve a testable performance gain under link-level testing channel conditions compared to non-AI. * **Step 8**: Specify the decoder from Step 5 and the encoder from Step 7. |
| R4-2510989 | ZTE, Sanechips | Observation 1. Based on our simulation results, transformer model has a better performance compared to other models and MLP has a worst performance.  Observation 2. For N-to-1joint training and separate testing, the simulation results could be very similar with one-to-one joint training and testing.  Observation 3. It is feasible to train a common decoder and test it with different models of encoder.  Propose 1. Propose to assume eigenvectors as default format for training dataset. |
| R4-2511248 | Samsung | *Remaining simulation for low-complexity encoder training*  **Observation 1:** To evaluate the feasibility of the common decoder compatible with different encoder backbones/complexities, the performance achieved for CNN-based backbones should be considered as the baseline performance (SGCS-0 for joint training and SGCS-1 for separate training).  **Observation 2:** By comparing the separate training for encoder-2/3/4 (SGCS-2) and encoder-1 (SGCS-1), the following evaluation purposes should be focused:   * **Purpose-(1): the feasibility of future-proof common decoder**, i.e., whether it is feasible to derive a common decoder which can be compatible with encoders with *different/new backbones* which has not been considered in the training. In other words, Purpose-(1) guarantee the RAN4-specified common decoder is future-proof for unseen AI/ML model backbone. * **Purpose-(2): the feasibility of low-complexity encoder**, i.e., whether it is feasible to have a separate training to derive *low-complexity encoder(s)* which is compatible with known common decoder with more complex backbone structure. In other words, Purpose-(2) guarantee the UE vendors can derive models with low complexity for implementation flexibility.   **Observation 3:** For a given new encoder-x with different backbone from encoder/decoder-1, by evaluating the performances for separate training for encoder-x (SGCS-2), separate training for encoder-1 (SGCS-1), joint training for encoder-x (SGCS-3) and joint training for encoder-1 (SGCS-0), the following conclusions can be made:   |  |  |  |  | | --- | --- | --- | --- | | What we observed | What we can conclude | What we observed | What we can conclude | | SGCS-3 SGCS-0 | Encoder-x backbone is compatible with the common decoder backbone | SGCS-2 SGCS-1 | (1) Separate training is feasible for encoder-x backbone;  (2) Specified common decoder is feasible and future-proof for unseen low-complexity encoder. | | SGCS-2 < SGCS-1 | (1) Separate training is NOT feasible for encoder-x backbone, while it is to be study the encoder-structure-aware training is feasible for this case.  (2) Specified common decoder is NOT feasible/future-proof if similar observation for different encoder backbones. | | SGCS-3 < SGCS-0 | Encoder-x backbone is NOT compatible with the common decoder backbone | N/A | |   **Proposal 1:** Use the term “N-to-1 Multi-encoder-structure-aware training” to represent the procedure for SGCS-4.  **Observation 4:** On top of Observation 3, the evaluation of multi-encoder-structure-aware training (SGCS-4, Step-2), the highlighted further conclusions can be made:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | What we observed | What we can conclude | What we observed | What we can conclude | What we observed | What we can conclude | | SGCS-3 SGCS-0 | Encoder-x backbone is compatible with the common decoder backbone | SGCS-2 SGCS-1 | (1) Separate training is feasible for encoder-x backbone;  (2) Specified common decoder is feasible and future-proof for unseen low-complexity encoder. | N/A | | | SGCS-2 < SGCS-1 | (1) Separate training is NOT feasible for encoder-x backbone, while it is to be study the multi-encoder-structure-aware training is feasible for this case.  (2) Specified common decoder is NOT feasible/future-proof if similar observation for different encoder backbones. | SGCS-4 (step-2) SGCS-1 | Multi-encoder-structure-aware training is required to consider encoder-x structure | | SGCS-4 (step-2) < SGCS-1 | Multi-encoder-structure-aware training methodology is NOT feasible for this case | | SGCS-3 < SGCS-0 | Encoder-x backbone NOT compatible with the common decoder backbone | N/A | | | |   **Observation 5:** For joint-training of Encoder-1/2/3/4 with the agreed decoder backbone (by comparing SGCS-3 for Encoder-2/3/4 to SGCS-0 for Encoder-1),   * The losses for Encoder-2, Encoder-3 and Encoder-4 are -2.0661%, 2.3368% and 3.9470%, all being regarded as similar performance as SGCS-0.   **Observation 6:** For the evaluated very low-complexity Encoder-4 (MLP-based low-complexity encoder), the SGCS from joint-trained encoder-4 has worse performance compared to other evaluated encoders.  **Observation 7:** For separate training for Encoder-1/2/3/4 with Samsung’s frozen decoder, i.e., SGCS-1 for Encoder-1 and SGCS-2 for Encoder-2/3/4, by comparing to the counterpart from joint training, i.e., SGCS-0 for Encoder-1 and SGCS-3 for Encoder-1:   * The losses for separate trained Encoder-1, Encoder-2 and Encoder-3 are 0.0285%, 1.5915% and 0.4085% respectively, all being degraded at a low level and showing similar performances as joint trained ones respectively; * The loss for Encoder-4 is 5.7707% which show obvious performance degradation.   **Observation 8:** For separate training for Encoder-2/3/4 with Samsung’s frozen decoder, i.e., SGCS-2 for Encoder-2/3/4, by comparing the performance from the separate training for Encoder-1, i.e., SGCS-1:   * The losses for Encoder-2 and Encoder-3 are -0.4704% and 2.7081% respectively, being degraded at a low level and showing similar performances as Encoder-1. * The loss for Encoder-4 is 9.4641% which show obvious performance degradation compared to other encoders.   **Observation 9:** The evaluated very low-complexity Encoder-4 (MLP-based low-complexity encoder) lack enough model capacity to compress the input features into the latent representation that the common decoder requires.  **Proposal 2:** For the evaluated very low-complexity Encoder-4 (MLP-based low-complexity encoder), it should not be recommended as reference encoder selection if the multi-encoder-structure-aware training is not considered:   * The low performance reason can be further studied for (1) MLP backbone; (2) very low complexity; or both (1) and (2).   **Proposal 3:** For the evaluated low-complexity Encoder-2/3 (transformer-based and CNN-based):   * Separate training with Samsung’s frozen common decoder is feasible for encoder-2/3 backbone   **Proposal 4:** The frozen Samsung’s decoder is feasible and future-proof for unseen low-complexity encoder, which also demonstrate the methodology of separate training.  **Observation 10:** When conducting multi-encoder-structure-aware training, it is demonstrated that n2 = 0.05 achieves the best average testing SGCS performance.  **Observation 11:** For multi-encoder-structure-aware training for Encoder-1/2/3/4 (i.e., SGCS-4, Step-2), by comparing to Encoder-1’s separate training (i.e., SGCS-1),   * The losses for Encoder-1, Encoder-2, Encoder-3 and Encoder-4 are 0.0713%, -1.7531%, 1.7531% and 3.6488% respectively, all being degraded at a low level and showing similar performances as Encoder-1’s separate training.   **Observation 12:** With the N-to-1 multi-encoder-structure-aware training, the joint trained decoder can deliver a new latent representation that is compatible to the small model capacity of Encoder-4.  **Proposal 5:** The feasibility of a common decoder for future-proof compatible to encoders with different backbones and with different complexities can be confirmed, by considering:   * Minimum complexities of the different encoder backbones should be guaranteed for the specified common decoder, if 1-to-1 separate training is used. * N-to-1 multi-encoder-structure-aware training can help to improve particular low-complexity encoder backbone.   *Remaining issues for non-simulation*  **Proposal 6:** For Option 3, the encoder that corresponds to the standardized test decoder is stored within 3GPP:  **Proposal 7:** For Option 3, the encoder that corresponds to the standardized test decoder is stored within 3GPP:  - It should be based on 1-to-1 joint-training of this encoder and corresponding standardized test decoder.  **Observation 13:** For Option 3, for dataset used for training and testing, the detailed content to be stored, “whether channel, Eigenvectors or something else” should depends on the standardized use case in Rel-20. |
| R4-2511571 | Qualcomm | Observation 1: Option 3 (fully specified decoder) leads to good SGCS in, at least, the scenario where encoder is trained with the same dataset as the decoder.  Observation 2: Option 4 leads to good SGCS between company’s own encoder and nominal decoders of all other companies when all encoders and decoders are generated based on Samsung’s reference encoder and mixed dataset.  Observation 3: Performance of option 4b can significantly drop if a different dataset is used while training the nominal decoder and the test decoder.  Observation 4: Recently agreed RAN plenary requires RAN4 to define requirement for the fully specified test decoder, i.e., option 3.  Observation 5: Both option 3 and option 4 are quite similar in terms of capturing or specifying a reference encoder, a reference decoder and a training dataset.  • Note: Option 4 allows test vendor to implement test decoder with different parameters or with a partially different structure  Observation 6: In option 4, UE’s failure during a test could stem from UE issue or from a potential TE issue where TE’s selected parameters for the test vendor are not compatible for the reference encoder and decoder  Observation 7: RAN4 has not calibrated system level channel for PMI testing. RAN4 has, so far, only calibrated TDL based channel models for PMI testing.  Observation 8: RAN plenary has also agreed to use both the structure and parameters of fully defined encoder/decoder in option C and just the structure in option 4a-1.  Observation 9: RAN1 considered transformer-based encoder structure of Figure 2 as an example in its study of feasibility and scalability.  Observation 10: Transformer-based encoder-decoder architecture can achieve similar SGCS as CNN based encoder-decoder architecture.  • Sequential training of transformer-based encoder and CNN decoder generates same SGCS as CNN based encoder-decoder.  Observation 11: The required number of flops for the considered transformer-based architecture is much smaller compared to the CNN based architecture that RAN4 is currently considering in study item.  Proposal 1: RAN4 follows the guidance of the WID of Rel-20 AI/ML and focuses only on fully specified decoder of option 3 during the WI.  Proposal 2: RAN4 selects TDL channel to test two-sided CSI feedback mechanism of Rel-20.  Proposal 3: RAN4 discusses which of following channels to consider in the training data set to design the fully defined/specified reference model of “Direction C”.  • Link level channel (e.g. TDL)  • System level channel  • Generalized dataset containing dataset of system level plus link level channel  Proposal 4: Select transformer as the backbone structure to define the standardized test structure of RAN4, and the structure of “Dir C” and “Direction A, sub-option 3a-1” of RAN1 during the work item. |
|  |  |  |
|  |  |  |

## Open issues summary

### Sub-topic 2-1: 2-sided CSI compression

**Issue 2-1: Lower complexity encoder simulations**

WF from RAN4#115:

**Further simulations (not agreement)**:

In order to further elaborate option 3 track 2 considering lower complexity encoders, interested companies may contribute simulations to RAN4#116. The following guidance, and also the guidance in R4-2508084 is provided and recommended for companies to follow in order to obtained aligned simulations. For option 3, track 2 consider the following investigations for lower complexity encoder structures.

* Performance assuming training and testing with the Samsung frozen decoder
* Performance achieved with a jointly trained low complexity encoder and decoder
* Performance achieved when a frozen decoder is trained based on all of the encoder structures (i.e. different encoder complexity)
* Consider transformer, CNN, MLP

A good set of results and observations has been provided by different companies. The aim in this discussion should be to try to draw out what observations may be commonly understood from the results.

* Proposals
  + Performance with agreed CNN encoder structure should be seen as a baseline (Samsung observation 1)
  + Differentiate between compatibility to different backbones and compatibility to different complexity levels (Samsung observation 2)
  + Use the term “N-to-1 Multi-encoder structure aware training” for SGCS-4 (Samsung proposal 1)
  + If 1-1 joint trained, the MLP shows significantly lower performance than CNN and transformer (CATT observation 1, Samsung observation 6)
  + SGCS loss is observed when low complexity encoders are separately trained with the fixed CNN decoder (Huawei observation 1)
  + If operated with the Samsung decoder, MLP shows significantly lower performance than CNN and transformer (CATT observation 1, Mediatek observation 4 for mixed dataset, Oppo observation 3, Samsung observation 7, 8)
  + Transformer shows the best performance (vivo observation 2, ZTE observation 1)
  + MLP shows worst performance (ZTE observation 1, Samsung observation 9; due to not enough model capacity)
    - Do not consider MLP if there is no multi-encoder aware training (Samsung proposal 2)
  + CNN and transformer can interoperate with the frozen Samsung decoder with reasonable performance (CATT observation 1, MEdiatek slightly worse performance observation 2/3 for mixed dataset)
  + The Samsung frozen decoder can interoperate with CNN and transformer encoders with separate training (Samsung proposal 3). This means that the frozen decoder trained on CNN can be seen as future proof (Samsung proposal 4)
  + MLP and transformer encoders trained with the frozen encoder with company dataset do not perform well (Mediatek observation 5)
  + With N-1 joint training, the CNN and transformer have comparable performance with the frozen decoder. MLP performs slightly worse, but there is better generalization to different structures (CATT observations 2/3, CMCC, oppo observation 3, Samsung observation 12, proposal 5)
  + N-1 training of the encoders and decoder gives better performance than training of the encoders with the fixed decoder (Huawei observation 2)
  + N-1 training of the encoders and decoder gives slightly worse results than 1-1 training of encoder(decoders (Huawei observation 3)
  + N-1 joint training of the decoder gives better results than 1-1 joint training with each of the encoders (Ericsson observation 1)
  + Encoders trained with a frozen N-1 decoder perform better than 1-1 joint trained encoders (Ericson observation 2)
  + SGCS is lower if lower complexity CNN encoders are used with N-1 joint training to create the decoder (Ericsson observations 3-4)
  + If N-1 training is used then one decoder can work with different encoder structures (vivo observation 2, oppo observation 3)
  + It is difficult to conclude that there is a benefit for training the decoder with multiple encoder structures or backbones (Ericsson observation 5)
  + 1-1 joint training and separate training have similar results (ZTE observation 2)
  + Derive the test decoder with multiple low complexity encoders with different background -8oppo proposal 3, Huawei proposal 2))
  + It is feasible to train a common decoder and test different encoders (ZTE observation 3)
  + For option 4, the performance difference between the CNN, MLP and transformer lower complexity encoders depends on the training and testing datasets (Mediatek observations 8-10)
  + Use a high complexity decoder for the test model; it does not make much difference whether trained with a single CNN or multiple encoders and will provide robustness for different encoder backbones (Nokia observations and proposal 1)
  + For the SGCS4 step 1, there is a dependency on the choice of the anchor and the performance of the other encoders. There is a need in training to report learning curve stability for each encoder and avoid bias. Discuss how to ensure fairness to encoders (KTL observations and proposals 1-2).
* Recommended WF
  + It is important to differentiate between the impacts of different encoder backbones and the impacts of different complexity levels
  + 1-1 joint training:
    - For 1-1 joint training, some, although not all, companies observed that the MLP encoder (Encoder 4) had a significant loss compared to the other encoders.
  + Training based on the Samsung frozen decoder
    - When trained using the Samsung Frozen decoder, many, although not all, companies observed that the MLP encoder had a significant loss compared to the other encoders
    - One company saw SGCS loss for all of the low complexity encoders (Transformer, CNN and MLP) when trained with the Samsung frozen encoder.
    - Based on the majority company results, it might be concluded that the lower complexity transformer and CNN encoders are compatible with the Samsung frozen decoder (which was trained using the agreed “high complexity” encoder), but the MLP may not be.
  + Training based on N-1 decoder
    - It is proposed to name training based on the N-1 decoder as “N-to-1 Multi-encoder structure aware training”
    - With N-to-1 Multi-encoder structure aware training, most companies observed that the performance of the MLP low complexity encoder could improve relative to the other encoders (depending on company, possibly still a slight loss)
    - One company observed that if lower complexity CNN encoders are used with N-1 joint training then the SGCS becomes lower for the low complexity CNN encoders.
    - One company observed that the choice of anchor and method used to train the N-to-1 Multi-encoder structure aware training may impact the performance and capability with different encoders.
  + Summary conclusion:
    - If a high complexity test decoder is used, it seems to be compatible with different encoder backbones (at least CNN and transformer)
    - The benefits of N-to-1 Multi-encoder structure aware training are not fully clear, however it may improve the performance of the lowest complexity MLP encoder
    - If the encoder complexity is not sufficient, then the performance will not be sufficient. However, to some extent a low complexity encoder may be compensated by a high complexity decoder.

**Issue 2-2: Field observations**

* Proposals
  + Based on field observations of 2 cells, the reference model performs the same or worse than eType II
    - The model can be refined
    - Refining the decoder gives better performance than refining the encoder
    - Refining both encoder and decoder can cause performance loss due to mismatch
    - (Vivo observation 6-7)
* Recommended WF
  + Discuss whether or not these observations should be captured in the TR

**Issue 2-3: Encoder / Decoder selection criteria**

* Proposals
  + Achievable performance, complexity, robustness (NTU proposal 1)
    - Robustness is investigated by comparing all decoder and encoder pair SGCS (or performance metric) (NTU proposal 2)
    - The above plus consider different UE capabilities (Oppo proposal 1)
  + For option 4, set a threshold for the difference between the reference decoder output and test decoder output, or even require the test decoder output compared to the text decoder output to be the same as the difference between the reference decoder output and the test ground truth (NTU proposal 3)
* Recommended WF
  + Achievable performance, complexity and robustness in different conditions need to be taken into account in selecting the model.
  + Issue 2-4 discusses the need for multiple models or not
  + For option 4, there is a need to set a “performance requirement” on the TE decoder to show that the performance compared to the reference decoder is satisfactory.

**Issue 2-4: Number of test decoders and generalization of performance**

* Proposals
  + Aim for one test decoder and reference encoder that can work over a wide enough range of conditions (Apple proposal 1)
  + Consider a frozen “backbone” plus scenario specific adapters if needed for some scenarios (Apple proposal 2)
    - Adapters created using scenario specific datasets and assumptions (Apple proposal 3)
  + Create a framework to compare single model vs per scenario model vs single model with backbone and investigate performance across scenarios to determine how many models and/or adaptors needed (Apple proposal 4)
  + Aim for one reference encoder per test decoder (Apple proposal 5, Nokia proposal 4)
  + If there are performance tiers with different encoders, determine a metric to define the tiers (Apple proposal 6)
* Recommended WF
  + An important issue to discuss is whether to allow for several different performance requirements depending on the encoder complexity, or a single performance requirement that each UE vendor meets with an encoder design of their choice.
    - Recommendation: Agree a minimum performance requirement level per set of side conditions (not multiple levels depending on encoder type)
  + Whether a single or multiple test decoders is needed will need more elaboration during the WI phase
  + Discuss further the proposal to assume a single model with “Adaptors”
    - More description of adaptors useful, and on the advantage of N “model + adaptors” compared to N models.

**Issue 2-5: Reference encoder**

**Agreements from RAN4#114bis**

* The encoder that corresponds to the standardized test decoder is stored within 3GPP
  + TBC whether in a TR, or in any TS, whether in RAN4 or RAN5, or whether simply kept in a 3GPP database (with no reference from TR or TS)
  + FFS whether several encoders would be captured (e.g. due to different complexity)

**New agreement from RAN4#115:**

The training dataset used to train the test decoder in 3GPP should be captured

* FFS whether channel, Eigenvectors or something else

**Agreements from RAN4#114bis**

* Dataset is specified for option 4a.
* Assume reference decoder is captured.

**New agreements from RAN4#115:**

* Reference encoder should be captured
* FFS whether the labelled dataset is based on Eigenvector, channel estimation or something else
* The study has only considered the case of a common assumption on model structure for the “own” test decoder. This corresponds to standardized model structure.
  + Structure refers to backbone, numbers of layers, type of layers and all description of the model, but not the parameters.
  + Interoperability in the case that the structure of the TE decoder is not specified has not been investigated

**Agreements from RAN4#114bis**

* Dataset is specified for option 4a.
* FFS on dataset is captured for option 4b.
* Assume reference decoder is captured.
* FFS on reference encoder for option 4a
* Test decoder structure is (i) fully specified or (ii) partially specified
* NOTE: RAN4 understands that further decision on option 4a/4b and option 3 will be made in the future WI phase.

**New agreements from RAN4#115:**

* Specify dataset to train frozen encoder and decoder
  + - TE vendor shall use the dataset to train the test decoder
    - This is the same dataset as used to train the frozen encoder
    - FFS whether the datasets consist of Eigenvector, channel or something else
* The study has only considered the case of a common assumption on model structure for the “own” test decoder. This corresponds to standardized model structure.
  + Structure refers to backbone, numbers of layers, type of layers and all description of the model, but not the parameters.
  + Interoperability in the case that the structure of the TE decoder is not specified has not been investigated
* Proposals
  + Specify a reference encoder (Nokia proposals 2-3)
  + Specify a reference encoder in the TS (Vivo proposal 1)
  + RAN4 to develop the reference encoder (vivo proposal 2)
  + Store the encoder in 3GPP. It should be 1-1 joint trained with the test decoder (Samsung proposals 6-7)
* Recommended WF
  + For the WI, RAN4 should develop the reference encoder (structure and parameters). The reference encoder should have been joint trained with the test decoder.

**Issue 2-6: Option 3 vs option 4**

Agreements from RAN4#115:

At the current moment in time, neither option 3 track 2, nor option 4a or 4b should be ruled out for a WI. Option 3 track 1 can be ruled out.

Option 3 and 4 are similar:

* An encoder, decoder and training dataset are always captured
* The difference is that for option 3, the TE vendor is mandated to use the standardized test decoder. In option 4 they could choose to use the reference test decoder or to create their own based on the standardized information.
* Proposals
  + Merge 4a/4b. Assume option 3 as baseline, check further option 4b with a defined criteria when to stop investigating; for example no loss of interoperability (Apple proposal 7/8)
    - There may be a transition from option 3 to option 4 in future releases (Apple proposal 8)
  + Options 3 and 4 are similar in terms of what is specified; the difference is the flexibility for the TE vendor. A problem for option 4 is that it may not be clear if a UE fail is down to the UE or the TE (Qualcomm observations 5-6)
  + Option 3 and option 4 provide similar SGCS results considering multiple (7) decoders (Mediatek observations 6-7)
  + Option 3 and option 4 work as long as the same dataset is used for encoder and decoder training for option 4 (Qualcomm observations 1-2)
  + Option 4 does not work if a different daset is used to train the test decoder compared to the nominal decoder (Qualcomm observation 3)
  + An option is deasible if (1) provides sufficient performance (e.g., ±5% from the SGCS obtained with Rel-16 eType2 codebook), and (2) can be standardized (e.g., based on the agreements about what to capture/specify for each option). (Nokia proposal 6)
  + The WI will be based on option 3 (Qualcomm observation 4, proposal 1)
* Recommended WF
  + For options 3, 4a and 4b, what is captured in the specifications is similar; encoder model, decoder model and dataset. The difference is whether the TE vendor is mandated to use the test decoder or can train their own (Previous agreement)
  + Option 4 can work as long as the decoder structure is specified and the same dataset is used.
    - Some validation of the TE decoder may be needed
  + The Rel-20 WID already assumes option 3

**Issue 2-7: Next steps for the Rel-20 WI**

* Proposals
  + Discuss new encoder/decoder architecture for Rel-20 WI (Ericsson proposal 5)
  + RAN4 to develop test decoder, encoder and structure for RAN4 and for direction C and A sub-option 3a-1 (Vivo proposal 3)
    - Next step is to align on the model structure and scalability based on Rel-19 RAN1 and RAN4 conclusions (Vivo proposal 3)
  + At least transformer is considered as the backbone for direction C and direction A 3a-1 (Vivo proposal 4, Qualcomm observation 9 and proposal 4)
  + RAN4 to take the most straightforward and suitable of the RAN1 scalability alternatives (Vivo proposal 5)
  + It would be high workload to consider all scalability alternatives from RAN1. Select one alternative as a baseline proposal (Huawei observations 4-6, proposal 3)
    - The baseline can be as follows (Huawei proposal 4)

|  |  |
| --- | --- |
|  | Scalability solution |
| Choice of token/feature dimension (if Transformer is agreed) | * Alt 1: Use subband as the token dimension and Tx port as a feature dimension   + The number of tokens varies with the number of subbands. |
| Scalability over the feature dimension | * Alt2: a common embedding layer with padding (e.g., zero-padding or other techniques for padding values) |
| Scalability over the token dimension | * Alt2: Padding at the input |
| Scalability over payload configurations | * Alt2: truncation/masking of the output linear layer output |

* + The following steps are proposed for the WI phase (Vivo proposal 6):
    - Step 0: Evaluation assumptions and mixed dataset based on Rel-19 can be reused.
    - Step 1: Align on the model backbone, model hyperparameters and the scalability solution.
    - Step 1-1: Proposed model structures (model backbone and model hyperparameters) are brought by companies, as well as proposed scalability solution.
    - Step 1-2: Check on performance alignment -> see simulation results from contributing companies.
    - Step 1-3: Share models (encoder or decoder or both)/datasets (training/testing/inference), if needed.
    - Step 1-4: Select one or more model structures.
    - Step 1-5: Select one or more scalability solutions. This step can be done in parallel with Step 1-4.
    - Step 2: Specify the model for the following work in WID. Where and how to specify may need further discussion.
    - Fully defined/specified reference encoder for testing decoder, for both 3a-1 and direction C, and
    - Fully specified test decoder.
  + The following steps are proposed for the WI phase (Huawei proposal 5)
* **Step 1**: Align on configurations for mixed dataset construction, taking scalability over Tx port number and subband size into account.
* **Step 2**: Align on the model backbones and model structures based on companies’ contributions with an aligned scalability method, where model complexity is limited to a predefined range. One model structure is selected for decoder, while one or more model structures are selected for encoder.
  + The baseline scalability method is as follows.
* **Step 3**: Align on training procedures (e.g., joint training and separate training, N-to-1 training if several encoders are used), hyperparameters and loss function definition.
* **Step 4**: Share encoders and decoders and cross-check.
* **Step 5**: Select one decoder with medium performance, that can achieve a testable performance gain under link-level testing channel conditions compared to non-AI.
  + If no decoders meet the requirement, go back to Step 2 to reselect the model structures.
* **Step 6**: Retrain encoders with the selected and fixed decoder. Share the encoder and cross-check.
  + If performance is not aligned, go back to Step 5 to reselect the fixed decoder.
* **Step 7**: Select one encoder from the retrained encoders which can achieve a testable performance gain under link-level testing channel conditions compared to non-AI.
* **Step 8**: Specify the decoder from Step 5 and the encoder from Step 7.
  + Use the mixed dataset for reference model training based on SLS and LLS for the RAN4 performance requirements (Vivo proposal 7)
  + The encoder/decoder model pairs for RAN4 requirements and RAN1 reference models may not be the same. RAN1 models will be related to testing requirements and RAN1 commercial deployments. RAN4 to develop models for both testing and for commercial deployments and evaluate whether RAN1s needs are met (Oppo observation 1, proposal 2)
  + RAN4 needs to discuss how to capture the model structures and the model parameters/weights in the 3GPP specifications (Ericsson proposals 1-2)
  + RAN4 needs to discuss the test metric for 2-sided CSI compression (Ericsson proposal 4)
  + RAN4 considers TDL for testing. Further discuss TDL/system level/mixture for training (Qualcomm proposal 3)
* Recommended WF
  + The following is needed for the WI:
    - Discuss and align on the dataset assumptions and dataset
    - Discuss and align on assumptions for the test decoder and reference encoder structures
      * Also may need to consider how many test decoders and encoders
    - Discuss and align on the scalability alternatives
      * Can the following be taken as a baseline ?

|  |  |
| --- | --- |
|  | Scalability solution |
| Choice of token/feature dimension (if Transformer is agreed) | * Alt 1: Use subband as the token dimension and Tx port as a feature dimension   + The number of tokens varies with the number of subbands. |
| Scalability over the feature dimension | * Alt2: a common embedding layer with padding (e.g., zero-padding or other techniques for padding values) |
| Scalability over the token dimension | * Alt2: Padding at the input |
| Scalability over payload configurations | * Alt2: truncation/masking of the output linear layer output |

* Discuss and agree on expectations on whether the decoder/encoder for RAN4 are the same as for option C
* Develop a means to capture model structures and model parameters in 3GPP
* Discuss and agree the test metric for the performance requirements.
* Discuss and agree channel model for the requirements.

**Issue 2-8: Dataset for WI**

* Proposals
  + RAN4 shall define a simulation framework covering frequency bands, environments, mobility, antenna setups, and SNR distributions. All companies must follow these simulation agreements, with dataset alignment verified via metrics like power spectral entropy (PSE). (Apple proposal 9)
  + New datasets need to be introduced to ensure model generality (Ericson proposal 6)
  + The dataset that is captured in the WI should be raw channel (Nokia proposal 5)
  + The dataset that is captured in the WI should be Eigenvectors (ZTE proposal 1)
  + The dataset format needs to be revisited for Rel-20 (Ericsson proposal 3, Samsung observation 13)
* Recommended WF
  + Discuss the following:
    - Eigenvectors or raw channel (or both) ?
    - What it is important to capture within the dataset
    - PSE or similar metrics for checking alignment.

### Sub-topic 2-2: TP to TR and simulation results collection

**Issue 2-9: Text proposal for the TR**

* Proposals
  + A text proposal for the TR is provided in R4-2510241
  + Offline discussion is requested to refine the TP as needed and prepare for approval.

**Issue 2-10: Simulation results collection**

* Proposals
  + A proposal for simulation results collection is provided in R4-2510880
  + Offline discussion is requested to refine the proposal as needed and prepare for approval.