**3GPP TSG- RAN WG1 Meeting #112bis-e R1-** **230xxxx**

**e-Meeting, April 17th – 26th, 2023**

Agenda Item: 9.2.2.2

Source: Moderator (Apple)

Title: Summary #1 on other aspects of AI/ML for CSI enhancement

# Introduction

This paper summarizes the discussion for agenda item 9.2.2.2.

## Contact information

Please provide your contact information.

|  |  |  |
| --- | --- | --- |
| Company | Name | Email |
| Moderator (Apple) | Huaning Niu | huaning\_niu@apple.com |
| NTT DOCOMO | Haruhi Echigo | haruhi.echigo.fw@nttdocomo.com |
| Qualcomm | Jay Kumar Sundararajan | jsundara@qti.qualcomm.com |
| NEC | Zhen He  Caroline Liang | he\_zhen@nec.cn  caroline.liang@emea.nec.com |
| CATT | Yongqiang Fei | feiyongqiang@catt.cn |
| LG Electronics | Haewook Park | haewook.park@lge.com |
| Spreadtrum | Hualei Wang | Hualei.wang@unisoc.com |
| Lenovo | Ahmed Hindy | ahmedhindy@motorola.com |
| Samsung | Caleb Lo | [caleb.lo@samsung.com](mailto:caleb.lo@samsung.com) |
| CMCC | Yuhua Cao | caoyuhua@chinamobile.com |
| MediaTek | Gyu Bum Kyung  Pedram Kheirkhah Sangdeh | [Gyubum.kyung@mediatek.com](mailto:Gyubum.kyung@mediatek.com)  [Pedram.kheirkhah@mediatek.com](mailto:Pedram.kheirkhah@mediatek.com) |
| Panasonic | Tetsuya Yamamoto | yamamoto.tetsuya001@jp.panasonic.com |
| Xiaomi | Zhengxuan Liu | liuzhengxuan@xiaomi.com |
| Sony | Hiroki Matsuda | hiroki.matsuda@sony.com |
| Sony | Sam Atungsiri | sam.atungsiri@sony.com |
| Intel | Victor Sergeev | victor.sergeev@intel.com |
| AT&T | Isfar Tariq  Salam Akoum | [Isfar.tariq@att.com](mailto:Isfar.tariq@att.com)  [Salam.akoum@att.com](mailto:Salam.akoum@att.com) |
| OPPO | Wenqiang Tian | tianwenqiang@oppo.com |
| Google | Yushu Zhang | yushuzhang@google.com |
| Huawei, HiSilicon | Yuan Li  Keyvan Zarifi | [liyuan3@huawei.com](mailto:liyuan3@huawei.com)  Keyvan.Zarifi@huawei.com |
| Fujitsu | Xin Wang  Qun Zhang | wangxin@fujitsu.com  zhangqun@fujitsu.com |
| CAICT | Xiaofeng Liu | Liuxiaofeng1@caict.ac.cn |
| Futurewei | Baoling Sheen | bsheen@futurewei.com |
| NVIDIA | Xingqin Lin | xingqinl@nvidia.com |
| InterDigital | Moon-il Lee | [Moonil.lee@interdigital.com](mailto:Moonil.lee@interdigital.com) |

# Potential specification impact for CSI compression with two-sided model

## Training collaboration

Three types of training collaboration were agreed in RAN1 110. Following table summarize company’s proposals and observations related to each type of training collaboration.

|  |  |
| --- | --- |
| **Company** | **View** |
| Huawei | Observation 4: For CSI compression with two-sided model, training Type 1 may suffer software/hardware compatibility issue, and the following restrictions/issues may need to be considered to relieve the compatibility issue:   * Network may have to interoperate with various UE vendors/UE versions to dedicatedly train the CSI generation part for UE, which harms the engineering isolation. * Network, in particular gNB, may have to maintain/store multiple CSI generation parts trained for different UE vendors/UE versions. * Network vendor may not freely develop the CSI generation part for UE, which may restrict the pairing with the CSI reconstruction part and thereby result in sub-optimal performance.   Observation 5: For training Type 1 of CSI compression, compared with joint training at Network side, performing joint training at UE side and delivering the model to the Network incurs extra challenges for Network due to the following reasons:   * Inconvenience of training cell/scenario specific models. * Inflexible model update. * Burden of inference/storing/running multiple Network part models at gNB delivered from different UE vendors/UE versions.   Proposal 10: For training Type 1 of CSI compression, deprioritize the mode of joint training at UE side and delivering the model to the Network.  Observation 6: For training Type 2 of CSI compression and model updating, it relies on complex design to support real-time interaction of FP/BP iterations between Network side and UE side, which causes strong challenges to engineering isolation especially for the case of multi-Network vendors to multi-UE vendors.  Observation 7: For training Type 3 of CSI compression,   * The shared dataset is constituted by the CSI-related data which may be irrelevant with the user privacy (e.g., user position, etc.). * The dataset sharing/delivery can be performed under the contract agreement between the Network vendors/MNOs/UE vendors to mitigate the data ownership problem.   Observation 8: For training Type 3 of CSI compression, compared with NW first training, performing UE first training incurs extra challenges for Network due to the following reasons:   * Inconvenience of training cell/scenario specific models. * Inflexible model update. * Burden of maintaining/storing multiple Network part models at gNB to pair with multiple UE vendors/UE versions. |
| ZTE | Proposal 1: Prioritize Type 1 joint training at NW side for further study and model transfer/delivery can be further discussed in agenda item 9.2.1.  Proposal 2: For training Type 3, prioritize NW-first training as a starting point for further study.  Proposal 3: For training Type 3, further study potential specification impact on the dataset used for the model training at the other side/entity.  Proposal 4: Conclude the pros and cons of different training collaboration types and prioritize training collaboration types in RAN1#112bis-e meeting. |
| OPPO | Proposal 1: In training collaboration type 3,   * For NW first training, NW needs to be able to provide UE with training data sets that meet different requirements, e.g. on model performance, transmission cost, data characteristics and CSI input types * For UE first training, UE needs to be able to provide NW with training data sets that meet different requirements, e.g. on model performance, transmission cost and data characteristics   Observation 1: For training collaboration types 1 and type 3, the conclusions for most of the questions listed below are the same, but the implementation methods are different. These two training collaboration types are not exclusive, and both can be considered in subsequent research. |
| vivo | 1. Pros/cons for training collaboration type 1: 2. Pros: Optimal performance 3. Pros: Provide highest flexibility in developing scenario-/configuration-/site-specific models via model transfer and model updating 4. Cons: Model proprietary could not be kept during model transfer. However, if trivial models are used, model proprietaries issue does not exist. 5. Cons: Require UEs to report the supported model design to develop device-specific models 6. Pros/cons for training collaboration type 2: 7. Pros: Model proprietary could be kept. However, if trivial models are used, model proprietaries issue does not exist. 8. Pros: Support device-specific models without the need to share model information to other entities 9. Cons: Need to share real-time information on forward /backward propagation result and label data. The overhead is very high to achieve near-optimal performances. 10. Cons: Lower flexibility to support cell/site/scenario/configuration specific model. Consequently, both sides need to train and store a large number of models to adapt to various scenarios/configurations 11. Pros/cons for training collaboration type 3: 12. Pros: Model proprietary could be kept. However, if trivial models are used, model proprietaries issue does not exist, 13. Pros/Cons: Support device-specific models without the need to share model information to other entities, but device-specific data distribution may not be supported. 14. Cons: Need to share information on dataset. May have risk in disclosing data from one user to another one. 15. Cons: Performance will degrade if shared dataset is insufficient or model structures are not aligned. 16. Cons: lower flexibility to support cell/site/scenario/configuration specific model. Consequently, both sides need to train and store a large number of models to adapt to various scenarios/configurations 17. The Extendibility issues (including training new UE-side model compatible with NW-side model in use and the support of “one-to-multi”/“multiple-to-one” configuration) could be addressed via proper training strategies for all training collaborations. |
| Spreadtrum | Proposal 1: Legacy CSI framework can be reused for the sub use case - Spatial-frequency domain CSI compression. Additional enhancement can be considered.  Proposal 2: To facilitate the discussion, views on Pros and Cons of all of Training types are needed to be aligned. What shown in Table 1 can be considered. |
| Nokia | Proposal 13: RAN1 shall investigate the appropriate dataset sharing without disclosing the mapping from (quantized) latent representation to the codeword.  Proposal 14: RAN1 may study the performance of the ML-based CSI-compression where the training datasets by the UE and gNB vendors are not matched. Also, it is necessary to study the case where the test scenario is not matched with the training scenarios of the encoder and decoder. |
| CATT | Observation 1: In CSI compression using two-sided model use case, for training collaboration Type 1, it has the following pros and cons:   * Pros:   + Optimal performance can be achieved;   + For joint training at UE side,     - Dataset sharing might not be needed.     - Maintaining/storing a single/unified model at UE side can be supported.   + For joint training at NW side,     - Cell/site/scenario specific model can be supported easily.     - Maintaining/storing a single/unified model at NW side can be supported. It is possible that UE does not need to maintaining/storing models for lots of gNBs. * Cons:   + Model transfer is needed.   + Model updating is lack of flexibility after deployment.   + For joint training at UE side,     - It is challenging for a UE to support cell/site/scenario specific model.     - A gNB has to maintain/store multiple models for multiple UEs.     - gNB specific optimization is not supported.   + For joint training at NW side,     - UE specific optimization is not supported.     Observation 2: In CSI compression using two-sided model use case, for training collaboration Type 2, it has the following pros and cons:   * Pros:   + gNB/device specific optimization is supported.   + Model transfer is not needed, which can keep model proprietary. * Cons:   + The latency on model training is large.   + There is heavy burden in air interface on real-time information exchange between NW side and UE side.   + Model updating is lack of flexibility after deployment.   + Further study is needed on maintaining/storing a single/unified model at both sides of network and UE.   + It is not easy to support cell/site/scenario/configuration specific model.   Observation 3: In CSI compression using two-sided model use case, for training collaboration Type 3, it has the following pros and cons:   * Pros:   + Model transfer is not needed.   + gNB/device specific optimization is supported.   + For sequential training starting with NW side,     - Good extendibility.     - Cell/site/scenario specific model can be supported by NW side easily.     - It is feasible to maintaining/storing a single/unified model at NW side.   + For sequential training starting with UE side,     - It is feasible to maintaining/storing a single/unified model at UE side.     - Model transfer is not needed, which can keep model proprietary. * Cons:   + Dataset transfer from the starting with side to the other side requires extra data transfer overhead.   + For sequential training starting with NW side,     - Further study is needed on the feasible of maintaining/storing a single/unified model at UE side to adapt to various NW sides.   + For sequential training starting with UE side,     - Bad extendibility.     - It is difficult to support cell/site/scenario specific model.     - Further study is needed on the feasible of maintaining/storing a single/unified model at NW side to adapt to various UEs. |
| Ericsson | Observation 1: Type 1 training collaboration seem not feasible in near term  Observation 2: Type 2 training collaboration seem not feasible in practice  Observation 3: Type 3 training collaboration where NW trains first may be a feasible approach to training  Observation 4: [Type 4] training collaboration where NW trains first, freeze the decoder and provide gradient transfer to UE side using API (for UE side training) may be a feasible approach to training  [Proposal 2 For CSI compression use case, it is a requirement that only training types and methods that enables a single decoder to be implemented in the network side is to be considered, irrespectively of the vendor origins of the connected UE devices and/or UE chipsets.](#_Toc131752939)  [Proposal 3 For CSI use case in this SI, down-prioritize any studies on model transfer unless it is the only solution that provides performance benefits over legacy CSI reporting](#_Toc131752940)  [Proposal 4 Define a training collaboration [Type 4], using a frozen decoder and gradient transfer using API, as a training method, according to the following description.](#_Toc131752941)  [Proposal 5 In the remaining work in this SI, for training collaborations that include the multi-vendor situation, assume [Type 4], NW first, frozen decoder and gradient transfer using API.](#_Toc131752942) |
| Xiaomi | Observation 1: NW may store and manage a lot of NW-sided part models for joint training of the two-sided model at UE side and UE-first separate training.  Proposal 1: Both joint training of two-sided model at NW side for Type 1 and NW-first separate training for Type 3 can be considered to train two-sided CSI compression AI/ML model. |
| Panasonic | Observation 10: Type 1 training involves the exchange of AI/ML model and then, requires some common AI/ML inference algorithm and common reference for model inference.  Observation 11: For Type 2 with offline training, if the consideration on the air interface specification impact on FP/BP interaction is not needed, there might be no Type 2 specific specification impact.  Observation 12: For Type 3 training collaboration with network-first training, at least the option that network generates training dataset to enable UE side supervised learning should be studied.  Observation 13: For Type 3, 3GPP may need to define some kind of requirement of CSI encoding by input and output relation, performance test or something else. The input for the training can be 3GPP specified channel model or field raw data. The output for the training can be something 3gpp defined output or network vendor specific information. The UE model performance can be checked by 3gpp specification or inter-operability test (IOT).  Observation 14: Type 3 with network-first separate training might be feasible options at least Re.18/19 timeline from standardization effort perspective. Type 1 with network sided training can be potential in the long-term. |
| CAICT | Proposal 1: Training type 1 at gNB/UE should be supported and training type 1 at gNB could be considered as starting point.  Proposal 2: Training type 3 should be supported for two-side model training. UE side AI/ML model information exchanging between UE and NW should be considered for dataset size control. |
| ETRI | Proposal 1: Consider further studies on performance improvement of AI models with training datasets from realistic channel estimation.  Observation 1: One possible performance improvement of AI models with training datasets from realistic channel estimation, is training a denoising function additionally.  Observation 2: To train the additional denoising function of the AI model for CSI compression, obtaining a training dataset with pairs can be required. |
| MediaTek | 1. For training type 2, discuss alignment of quantization/dequantization as well as format/precision of gradient vectors, latent vectors, and CSI samples. 2. For single-encoder multi-decoder setting in training type 2, UE should not break down the training session into multiple single-encoder single-decoder training sub-sessions. 3. In training type 2 for multi-encoder setting, if UE-specific datasets are used, the type of target CSI should be aligned among UE vendors. 4. Discuss feasibility of synchronization/alignment required for different update scheduling in training type 2. 5. If UE-specific datasets are used for multi-encoder training, consider sharing information on training-related parameters such as size of datasets, statistics of datasets, training loss, and update schedule. 6. Consider sharing information about encoders’/decoders’ architecture type and complexity from entities doing training first to other entities. 7. Prioritize the study of training-aware quantization methods 8. Study alignment requirement and influence of different training awareness techniques for enabling backpropagation between quantizers and dequantizers. 9. For training type 2, gNB should inform UE about the training awareness technique used for its dequantizer. |
| Apple | Proposal 1: Model update using training collaboration type 2 over 3GPP air interface incur high complexity and large overhead. It can be deprioritized for R18 study  Proposal 2: To facilitate future discussion on necessity and benefit of each training collaboration type,   * Further categorize the training collaboration type 1 as: 1a-training at UE side, 1b-training at NW side. * Further categorize the training collaboration type 3 as: 3a-UE first and 3b-NW first. |
| Lenovo | 1. Study the training collaboration types considering the communication overhead and/or the corresponding latency, based on whether the communication between the network and UE sides during model training and model adaptation occurs over the NR air interface or via proprietary signaling 2. Study the advantages/disadvantages of joint training at the UE side vs. joint training at the network side with Type 1 training collaboration 3. Study the performance of iterative separate training as one of the methods to improve the performance of sperate training when multiple vendors are involved in training on the two sides of communication 4. For FDD systems with network-based Type-1 model training as well as Type-3 training collaboration, signaling the CSI training data from the UE to the network is needed 5. Evaluate schemes related to transfer of CSI dataset for different stages of the LCM 6. Evaluate the following CSI training data signaling techniques:  * Alt1. Proprietary signaling via non-3GPP techniques * Alt2. Legacy CSI dataset feedback where the NR codebook-based CSI is utilized as CSI training data * Alt3. Explicit CSI-dataset feedback via enhanced 3GPP-based signaling of the CSI training data  1. Evaluate the following CSI training data formats:  * Alt-A. Legacy codebook-based dataset points generated via multiple occasions of NR codebook-based CSI feedback * Alt-B. High-resolution codebook-based dataset points generated via high-resolution variants of NR-based CSI codebooks * Alt-C. Floating point representation of raw CSI data, e.g., raw channel matrices or sets of channel eigenvectors |
| Qualcomm | Observation 1: For the AI/ML-based CSI feedback enhancement use case, the use of an AI/ML model for inference within a device would require prior offline device-specific optimization and testing.  Observation 2: Type 1 training with device-agnostic encoder would result in a UE-side model that:   * + is not optimized in a device-specific manner for the intended UE-side device,   + assumes a structure and input format that is not compatible with the UE-side implementation capabilities, and   + may have sub-optimal performance due to a discrepancy between the training and inference data distribution due to device-side variations.   Observation 3: Type 1 training performed on the NW-side with involvement of the UE-side vendor requires the UE-side to provide information (such as model structure, pre-processing, post-processing, datasets and ground truth) to the training entity to ensure that the trained models are suitable for inference.  Observation 4: For NW-side type 1 training with UE-side involvement, developing a new model for a new UE device type or vendor can result in a large engineering effort across multiple vendors.  Observation 5: It is feasible to train a two-sided AI/ML model using an offline Type 2 (multi-vendor) training approach with performance comparable to Type 1 training.  Observation 6: For type 2 training, developing a new model for a new UE device type or vendor can result in a large engineering effort across multiple vendors if the NW-side or UE-side use a common model for multiple models on the opposite side.  Observation 7: As compared to Type 2 training, the Type 3 offline training approach is more flexible as it does not require coordination during the training process.  Observation 8: For Type 3 separate training, the engineering effort of adding a new UE type or new UE-side vendor is contained and does not propagate to other vendors even if the NW-side or UE-side use a common model for multiple models on the opposite side.  Observation 9: For NW-first sequential training, the training based on gradient exchange provides several benefits in terms of flexibility in the input type, better alignment between the UE-side and NW-side model training, aligned dataset and avoiding disclosure of proprietary information.  Observation 10: It is feasible to train a common NW-side model that is compatible with multiple UE-side models using Type 2 or Type 3 training approach with performance comparable to Type 1 training.  Observation 11: Training type 1 (with device-specific encoder), training type 2 and training type 3 are applicable to both collaboration level y and level z.  *Proposal 1:* For data collection for model training, RAN1 should focus on what data should be collected. Mechanism for training data collection needs architectural considerations and should be handled by other working groups.  Proposal 2: For AI/ML-based CSI feedback using two-sided model, the procedure used to process the downlink measurements and derive the input to the UE-side model during inference should be left to UE implementation.  Proposal 3: While generating the training dataset, the target CSI corresponding to a downlink measurement should be derived by the UE side to reflect the UE processing during inference (e.g., channel estimation, eigen-vector derivation, etc.).  Proposal 4: Study assistance signalling for UE’s data collection in the form of a zone ID, scenario ID, and configuration ID.  Proposal 5: Model development and training options should consider the need for the UE-part of two-sided AI/ML models to be designed based on the UE capabilities and optimized in a device-specific manner.  Proposal 6: Model development and training options should strive for the principle of engineering isolation, i.e., confining engineering effort needed for a new chipset/UE development to the given chipset/UE vendor.  Proposal 7: Model development and training options need to consider whether the model is developed for common use across a group of UEs or is developed for an individual UE.  Proposal 8: Model development and training options need to consider feasibility of disclosing proprietary model information to the other side.  Proposal 9: For AI/ML-based CSI feedback enhancement use-case, take offline training as a starting point.  Proposal 10: Deprioritize Type 1 training with device-agnostic encoder in the R18 study.  Proposal 11: Adopt the following two-sided model development/training framework:   * Case 1: Initial (non-backward-compatible) development/training of “nominal encoder + nominal decoder”   + The use of the nominal encoder at the UE-side is not mandated     - If needed, UE-side may implement a different proprietary encoder based on this decoder using Case 2.     - As the encoders are only nominal, input used in the training process is only a nominal input. The actual input to the CSI encoders may be different and of proprietary choice.   + The use of the nominal decoder at the NW-side is not mandated     - If needed, NW-side may implement a different proprietary decoder based on this encoder using Case 3. * Case 2: Encoder development/training to be interoperable with existing decoders (e.g., encoders for new UEs or updating encoders for existing UEs):   + UE-side vendor trains new encoders based on the existing decoders.   + Infra vendor should make the existing decoders available (via either a run-time image or an API for training) for the encoder training. * Case 3: Decoder development/training to be interoperable with existing encoders (e.g., decoders for new cell sites or updating decoders for existing cell sites):   + Network-side vendor trains new decoders based on the existing encoders.   + FFS: Need for encoder availability for decoder training |
| AT&T | Proposal 1: In CSI compression using two-sided model use case with training collaboration type 1, further study potential specification impact on:   * Protocol and signalling mechanism to enable CSI compression specific model transfer.   Proposal 2: In CSI compression using two-sided model use case with training collaboration type 3, for sequential training, further study necessity, feasibility, and potential specification impact on:   * Training dataset and/or other information delivery from UE side to NW side for UE first training * Training dataset and/or other information delivery from NW side to UE side for NW first training * Data sample format/type and the dataset size * Quantization/de-quantization related information * Note: other aspects are not precluded. |
| NTT DOCOMO | Observation 1: The performance of joint training is the upper bound of sequential training.  Observation 2: Three type of training procedures provides the similar performance, when the pre-/post-processing is aligned.  Observation 3: Type 2 and type 3 training procedure requires large signalling overhead due to the dataset transfer or the exchange of forward/back propagation from one side to the other.  Observation 4: Type 1 training procedure can provide good performance and requires less overhead signalling. However, the feasibility of model transfer is questionable in terms of the proprietary and hardware aspects.  Proposal 1: Categorize type 3 training procedure (sequential training) as follows.   * Type 3-A: sequential training via the dataset delivery   + Step 1 Joint training at one side   + Step 2 Delivery of the dataset produced by trained encoder/decoder to the other side   + Step 3 Training encoder/decoder based on the delivered dataset within the other side * Type 3-B: sequential training via the gradient exchange   + Step 1: Joint training at one side   + Step 2: Share the common dataset for training at Step 3   + Step 3: Training encoder/decoder at the other side via FP/BP exchange with the frozen decoder/encoder   Proposal 2: Deprioritize type 2 training procedure even for the model update. |
| Samsung | Proposal 2-6: Deprioritize two-sided model training collaboration that requires extensive sharing of training, validation and testing datasets over the air-interface in this study item.  Proposal 2-7: Study the impact of the following factors on two-sided model development approaches:   * Requirements on privacy-sensitive dataset sharing * Scalability, i.e., whether the number of models one vendor should develop increases with the number of collaborating vendors * Whether two-sided model development approaches adhere to 3GPP’s open and fair framework   Proposal 2-8: For Type 3 training collaboration, study performance impact of training/testing an encoder with a reference decoder or dataset.  Proposal 2-9: For AI/ML based CSI compression sub-use case, study and verify model update of the encoder at the UE, where the gNB’s training strategy is not disclosed while transferring/configuring the AE. |

### Summary:

Metrics to facilitate pros/cons discussion of each training collaboration type were captured in RAN1 112. Companies have provided views on the pros/cons of each training collaboration type. For training type 1 and training type 3, some analysis further separates it into UE side/NW side, and UE first/NW first. With early agreement that type 2 training collaboration is implementation-based solution based on multi-vendor agreement, some companies did not include type 2 training collaboration in the analysis. R1-2302919 further categorize each training collaboration type into multiple sub-types for detailed analysis.

FL tries to include high level aspects submitted to get comprehensive view. The following table summarizes the discussion.

### ***Proposed observation 2-1-1:***

***In CSI compression using two-sided model use case, the following table capture the pros/cons of different offline training collaboration types:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Training types  Characteristics | Type 1 | | Type 2 | Type 3 | |
| NW-sided | UE-sided |  | NW first | UE first |
| Whether model can be kept proprietary | No | No | Yes | Yes | Yes |
| Whether require privacy-sensitive dataset sharing | No (Note 1) | No | No (Note 1) | No (Note 1) | No (Note 1) |
| Flexibility to support cell/site/scenario/configuration specific model | Yes | Yes. With assisted information signaling | Difficult | Semi-flexible. | Semi-flexible. With assisted information signaling |
| Whether gNB/device specific optimization is allowed | Restricted | Restricted | Yes | Yes | Yes |
| Model update flexibility after deployment | Flexible | Flexible | Not flexible | Semi-flexible | Semi-flexible |
| Feasibility of allowing UE side and NW side to develop/update models separately | Limited  (Note 2) | Limited  (Note 2) | Infeasible | Feasible | Feasible |
| Whether gNB can maintain/store a single/unified model | Yes | No | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 |
| Whether UE device can maintain/store a single/unified model | No | Yes | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 |
| Extendibility: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use | Limited  (Note 2) | Limited  (Note 2) | Limited | Support | Support |
| Whether training data distribution can match the inference device | Restricted | Yes | Restricted | Restricted | Yes |
| Software/hardware compatibility (Whether device capability can be considered for model development) | Limited | Limited | Compatible | Compatible | Compatible |
| Model performance based on evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 | Pending evaluation in 9.2.2.1 |

Note 1: Assume high accuracy PMI is not privacy sensitive data. FFS: other information such as channel matrix and assisted information.

Note 2: For example, after deploying Model 1 on the UE side, a new UE model can be obtained by using Model 1 as the teacher model and using knowledge distillation method.

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| **Google** | Support in principle. One minor thing could be that for the two rows: “Whether gNB can maintain/store a single/unified model” and “Whether UE device can maintain/store a single/unified model”, we think we should add “For a CSI report configuration”.  If the NW provides different CSI report configurations with different CSI-RS, e.g., different CSI-RSs with different ports, a single/unified model may not work. |
| NTT DOCOMO | We could not find why the model update after deployment is flexible only for type 1 model training. The proponent should clarify it. |
| CATT | Generally OK, but:  (1) What does ‘training data distribution can match the inference device’ mean, and why only UE-sided/UE-first training can do it? Does ‘inference device’ only mean UE-side here? Can proponent clarify?  (2) Should we consider a column for *parallel training* in Type 3 collaboration? |
| Huawei/HiSi | **For “Model update flexibility after deployment”**  – for Type 1-UE side, if the flexibility can only be achieved based on the assisted information to achieve, it should be changed as “conditional, with assisted information”, or directly say “Restricted”.  – for Type 3-UE first, similarly, changed as “conditionally semi-flexible, with assisted information”, or directly say “Restricted”.  **For “Whether training data distribution can match the inference device”**-for Type 1-NW side and Type 3-NW first, it should be symmetric with the impact of “Model update flexibility after deployment” to UE side/UE first, both either aligned as “conditional, with assisted information” or aligned as “Restricted”.  **For “Model update flexibility after deployment”** – for Type 1-NW side, gNB can grab data from all UEs in the cell to quickly generate the updated dataset, and train the model on gNB; for Type 1-UE side, a UE vendor can only rely on its own UEs which is only a portion of UEs in that cell; moreover, the data needs to be uploaded to the non-3GPP entity from UE device rather than training on UE. So Type 1-UE side is “less flexible than Type 1-NW side”.  **For “Whether gNB can maintain/store a single/unified model”** – For Type 1-NW side, due to the software/hardware compatibility issue at UE, different UE/chipset vendors may have different flavors on the model structure, so the gNB has to store multiple CSI generation parts from different UE vendors/versions. So it should be changed to “No”  **For “Whether UE device can maintain/store a single/unified model”** – for Type 1-NW side, Type 2, Type 3-NW first, and Type 3-UE first, we believe that the UE part models are maintained at the UE side non-3GPP entity rather than the UE device, so it is feasible to store only one encoder at the UE device. In that sense, all 5 modes should be “Yes” |
| vivo | We are generally ok with the table with some minor comment:  1. In “whether model can be kept proprietary in training type 3”, we believe “Yes” should be “Partially Yes”, as information on model structure may be required to disclose.  2. For note2, we understand that model 1 can also refer to a nominal model while the real deployed model can be developed based on the nominal model.  3. We want to know, to which training collaboration the method of freezing decoder and updating encoder belongs when FL summarizes the table? |
| Ericsson | Knowledge distillation method needs to be defined. Also, I think this is seen as a starting point for further discussion. We may update or fine tune the observations later. |
| Xiaomi | We are fine with proposal. But we prefer to discussing Type I-NW side and Type 3-NW first with high priority. |
| LG Electronics | Support in principle. For the rows with pending evaluation in 9.2.2.1, we can skip those rows. If relevant agreement/conclusion are made in 9.2.2.1, we can add that rows. |
| OPPO | **For “Flexibility to support cell/site/scenario/configuration specific model”**  - for type 3 NW side first training, we think the answer is also **“Yes”.** Because NW can collect cell/site/scenario/configuration specific data and train the NW model, and transmit related **cell/site/scenario/configuration specific data** (to train the CSI encoder) to UE to help finishing the UE side model training. It is necessary to clarify what restrictions exist if it is “Semi flexible”  **For “Model update flexibility after deployment”,**  - What causes “Semi flexible” in type 3 training? Adding a note to clarify the difference between “flexible” and “semi flexible” would be helpful, if needed.  And agree with Xiaomi, we should focus on type1 and type3 since we have concluded that “In CSI compression using two-sided model use case, training collaboration type 2 over the air interface for model training (not including model update) is deprioritized in R18 SI”. |
| CMCC | For “Model update flexibility after deployment”, not sure why type 1 training can be flexible while type 3 training is semi-flexible.  Both of them need some additional dataset, the difference between them is just type 1 training need model delivery while type 3 training need dataset delivery. |
| CAICT | Fine in general. |
| Lenovo | For “**Flexibility to support cell/site/scenario/configuration specific model**” for type-3 what do we mean by “Semi-flexible”? Similar to Type-1 the NW-first and UE-first can update their local model and then signal the other party with the required dataset for updating. Even in some cases, like freezed decoder, the UE side can update its encoder based on the local data. So, we suggest using “yes” in these cases as well  For “**Whether model can be kept proprietary**”  **@vivo**: for type-3 we can manage to keep even the model structure in private by iterative execution of training, please take a look at “iterative type-3” that we have discussed in our tdoc. So, we suggest to keep this part as “yes”  **For “Model update flexibility after deployment”**  We are not sure what is the meaning of “Semi-flexible” vs “flexible”? |
| NVIDIA | Support the observation in principle. |
| ZTE | Generally fine with the table. |
| Intel | For **For “Model update flexibility after deployment”**, our view is aligned with comment from Huawei/HiSi: some level optimization is still possible (based on assistance info e.g., UE capability). |

## Data collection

Following table summarize company’s proposals related to data collection.

|  |  |
| --- | --- |
| **Company** | **Key Proposals/Observations/Positions** |
| Huawei | Proposal 2: For the enhancement of CSI-RS configurations for Network/UE side data collection under CSI compression, separate CSI-RS resources/CSI reports can be adopted for generating ground-truth CSI labels (e.g., measured with higher power/density CSI-RS) and model inputs (e.g., measured with lower power/density CSI-RS) to support the super resolution of using low resolution input to infer high resolution output.  Observation 1: It is necessary to support real time UE report of the monitoring results to gNB to enable fast identification of network performance fluctuation/degradation and AI/ML model failure.   * E.g., in case of performance degradation, event-triggered monitoring window can be activated so that gNB can efficiently collect data and thereby quickly identify whether the degradation is due to the AI/ML model failure.   Observation 2: For Network side monitoring based on intermediate KPI, the reporting of the ground-truth CSI and the associated CSI report by the UE via L1 signaling has comparable overhead with L3 signaling but with much less latency which can enable fast identification of AI/ML model failure.  Proposal 3: For the container of Network side data collection under CSI compression,   * Both of L1 signaling and RRC signaling can be supported for model training. * At least L1 signaling should be supported for model monitoring to enable fast identification of AI/ML model failure.   Proposal 4: For data sample format of Network side data collection under CSI compression:   * Both of scalar quantization and codebook-based quantization can be supported for model training. * Codebook-based quantization should be supported for model monitoring.   Proposal 5: For the Network side data collection of ground-truth CSI, the number of ranks and the index(es) of layer(s) for the report of ground-truth CSI can be designated by the gNB rather than autonomously calculated and reported by UE.  Observation 3: The applicable cases for the categorization ID for assisting UE side data collection may need to be further clarified with respect to the following points:   * Generalized model can be trained at UE side over scenarios/antenna layouts. * UE can autonomously sense the scenario without being notified by gNB. * The categorization or granularity of the scenarios identified by Network may not match the categorization principle of the UE side.   Proposal 6: The data categorization ID, if justified, should be determined by Network side in forms of virtualized ID without specifying the physical meaning.   * The physical meaning of such ID and the granularity of such ID is up to Network implementation without being indicated to the UE side.   Proposal 7: In CSI compression using two-sided model with training collaboration Type 3, further study potential specification impact of dataset delivery over air-interface on the following aspects:   * Training dataset and/or other information delivery from UE side to Network side for UE first training. * Training dataset and/or other information delivery from Network side to UE side for NW first training. * The specification impact includes dataset ID, the size of the dataset, format of data sample, type(s) of the data sample, quantization/de-quantization related information etc.   Proposal 8: For the dataset delivery of CSI compression over air-interface, the following approaches can be considered to largely alleviate the overhead/power consumption of per UE:   * Quantization on the ground-truth CSI with high resolution quantization format, e.g., R16 Type II-like method with new parameters. * Network splits the overall dataset into massive subsets each with limited number of data samples (e.g., with comparable overhead as RRC signaling). The subsets can be separately sent to numerous UEs, and all subsets are associated with a common dataset ID for UE side combination.   Proposal 9: For the dataset delivery of CSI compression over air-interface, the dataset ID associated with the delivered dataset can be used to achieve the pairing of the Network part model and the UE part model.   * E.g., for NW first separate training, UEs receive the dataset associated with a dataset ID to perform the training; after the UE part model is trained, UE and gNB will use the dataset ID to achieve the pairing. |
| ZTE | Observation 1: When model training or monitoring is performed at network side, the overhead of the ground-truth label transmitted over the air-interface from UE to network is a huge concern if the ground-truth CSI is an ideal CSI (e.g., raw channels, eigenvectors).  Observation 2: The overhead of enhanced Type II CB (i.e., PC10) for one training sample increases by 50% compared with the maximal payload of Rel-16 TypeII CB (i.e., PC8) but keeps similar model performance as ideal CSI, which can be acceptable to be carried on UCI.  Proposal 5: For network side data collection, support to further study   * Enhanced Rel-16 TypeII codebook to get high-resolution CSI; * PHY signaling or RRC signaling to report the high-resolution CSI.   Proposal 6: To enable high-quality data collection, at least support:   * UE reports associated information to NW, e.g., SINR, CQI, positioning information * NW configures a threshold of data quality to UE and UE only reports the qualified data to NW   Observation 3: For Type 3 training collaboration of a two-sided model, common understanding on the dataset used for model training is necessary, which can facilitate the pairing of CSI generation part and CSI reconstruction part.  Observation 4: Dataset alignment between UE and network can be used for testing/monitoring the model/functionality performance.  Observation 5: Dataset ID can avoid sharing the proprietary information explicitly across vendors during data collection.  Proposal 7: Support to use dataset ID for collected data and for delivered dataset of Type 3 training collaboration. |
| vivo | 1. Meta information reporting for data collection should be studied to facilitate the development of scenario-/area-/configuration-specific models. 2. The necessity of reporting certain kind of meta information in data collection depends on model’s generalization ability on it. 3. Enhanced legacy codebook can be used for data collection (CSI measurement), and enhancements for different data collection purpose can be different 4. RAN1 could send LS to RAN2 to clarify the requirement of data collection in CSI compression (and other use cases). |
| Spreadtrum | Proposal 3: For AI/ML model training type 2 and type 3, data collection is needed.  Proposal 4: For AI/ML model training Type 1, data collection may be not needed to be specified other than assisted signalling, e.g, antenna layout for one CSI-RS resource  Proposal 5: Offline AI/ML model training is the first priority.  Proposal 6: If model transfer supported, data collection procedure is not needed.  Proposal 7: If model transfer not supported, for UE side, data collection procedure may be needed.  Proposal 8: If model transfer not supported, for NW side, data collection procedure may be needed or not depending on whether SRS can be utilized.  Observation 1: It may be not necessary to do data collection for model monitoring |
| Nokia | Proposal 9: In CSI compression using a two-sided model, consider the following for the data collection,   * Data collection shall be mainly focused on performance monitoring or model fine-tuning, and considerations on the data collection for model training shall not be the main focus. * UE-sided data collection,   + Existing CSI-RS configuration shall be used as the starting point for any form of data collection * NW-sided data collection, * Enhancement of CSI reporting to enable higher accuracy reporting * FFS: Assistance information reporting |
| CATT | Proposal 1: In CSI compression using two-sided model use case, focus on studying NW side data collection in Rel-18 SI.  Proposal 2: In CSI compression using two-sided model use case, for NW side data collection for model training, focus on studying CSI-RS measurement based data collection.  Observation 4: In CSI compression using two-sided model use case, for data collection for model training, enhancement on CSI-RS is not needed.  Proposal 3: In CSI compression using two-sided model use case, on ground-truth CSI reporting for NW side data collection for model training, study potential specification impact on the following schemes:   * Option 1: Ground-truth CSI samples are reported by physical layer signaling, with legacy CSI feedback framework reused; * Option 2: Ground-truth CSI samples are reported by RRC signaling, with a batch of ground-truth CSI samples reported together.   Proposal 4: In CSI compression using two-sided model use case, for NW side data collection for model training, collecting ground-truth data in type of precoding matrix is supported.  Proposal 5: In CSI compression using two-sided model use case, for NW side data collection for model training, codebook-based quantization for ground-truth data is with higher priority. |
| NEC | Proposal 2: Study the mechanism of obtaining RSs specific for data collection in model training, model update and model monitoring, e.g., explicit configuration, implicit acquirement. |
| Ericsson | [Proposal 1 For CSI compression use case, it is required that standardized procedures and associated data format for UE to gNB data collection of a high-resolution CSI (target CSI) is supported to enable model monitoring and to provide data for enabling decoder fine tuning.](#_Toc131752938)  [Proposal 6 For CSI use case in this SI, down-prioritize studies on model transfer](#_Toc131752943)  Observation 5: 3GPP specifications needs to support a mechanism to update/fine tune the decoder to consider implementation reality (e.g., UE and gNB RF and antennas at UE and gNB) and to ensure good generalization performance in scenarios not part of the pre-deployment training dataset  Observation 6: A Target CSI definition approach based on the eType-II framework, with more selected beams, taps, and coefficients compared to existing eType-II, and with finer resolution in the quantization of the coefficients has the potential to accurately describe the true Tx-eigenvector.  Observation 7: Specification of UE to network data collection of UE measurements of target CSI is motivated by both monitoring and decoder adaptation purposes  Observation 8: An initial estimate of the data size for collection in the CSI compression use case is in the range of 10k-40kbit. Further detailed studies are needed. |
| Xiaomi | Observation 2: For UE side and network side collection, it is not necessary to enhance CSI-RS resource since higher accuracy channel measurement can be obtained by current CSI-RS resource configuration.  Observation 3: For UE side and network side data collection, the data ID is not necessary to report as an assistance information considering that the data ID can be obtained in a proprietary way or network configuration.  Proposal 2: For network side data collection, cell-specific CSI-RS resource configuration can be supported to reduce configuration signalling overhead.  Proposal 3: The design trigger signalling between UE and network should be specified.  Proposal 4: The methods on overhead reduction for AI/ML model training should be studied for updating or monitoring AI/ML model. |
| Panasonic | Observation 1: In CSI compression using two-sided model use case, it is necessary to use the ground-truth CSI of realistic DL channel measured by UE and report to network.  Observation 2: Data collection for model training is not required to be real-time and then latency requirement can be larger.  Observation 3: At least for data collection for performance monitoring, in order to handle multiple UE vendors and/or UE models, the reporting of ground-truth CSI should be performed using 3GPP signaling to avoid the complexity of handling multiple formats.  Observation 4: Depending on the requirement of latency, grouped reporting could be realized through MAC-CE RRC or U-plane, and sample-by-sample reporting is better to be implemented via UCI.  Observation 5: On data sample type / format for ground-truth CSI reporting, high resolution codebook-based format e.g., legacy codebook (e.g., eType II codebook) with potential enhancements such as extend more configurations in some parameters, should be studied.  Observation 6: For network-side data collection, at least time stamps/cell ID and UE location should be considered as the assistance information.  Observation 7: For network-side data collection, the necessity and feasibility of UE reporting Rx antenna spacing and Rx RF gain imbalance to network should be studied.  Observation 8: For UE-side data collection, to identify the scenario / configuration in which the data is being collected, virtualized configuration ID should be studied as the assistance information.  Observation 9: If CSI-RS / SRS configurations in current NR specification is not sufficient for higher accuracy measurement, enhanced CSI-RS and/or SRS may be considered for the data collection. |
| LGE | Proposal #1: For UE side data collection, deprioritize discussions on enhancement on CSI-RS configuration.  Proposal #2: For NW side data collection, deprioritize discussions on latency requirement and enhancement on SRS and/or CSI-RS configuration. |
| MediaTek | 1. While gNB is main entity in establishing data collection procedure, UE should provide gNB with a range of possible options for configurations of the data collection procedure including but not limited to:  * Types of input CSI * Types of assistant information * Quantization parameters * Periodicity of data collection * Maximum amount of data collected per period  1. Discuss the quantization of data in the following aspects:  * Decisioning entity * Incorporation of non-quantized CSI for possible finetuning * Quantizable information (CSI samples and assistant information) * Configuration changes (per sample or per dataset) |
| Nvidia | Proposal 6: For AI/ML model training for CSI feedback enhancement, study potential specification impact related to training data type/size, training data source determination, and assistance signalling and procedure for training data collection.  Proposal 7: For AI/ML based CSI feedback, study potential specification impact related to assistance signalling and procedure for model configuration, model activation/deactivation, model recovery/termination, and model selection.  Proposal 8: For AI/ML based CSI feedback, study potential specification impact related to assistance signalling and procedure for model performance monitoring and model update/tuning.  Proposal 9: For AI/ML based CSI feedback, study potential specification impact related to report/feedback of model input for inference (e.g., quantization and feedback message size), type of model input, and model input acquisition and pre-processing.  Proposal 10: For AI/ML based CSI feedback, study potential specification impact related to report/feedback of model inference output (e.g., quantization and feedback message size) and post-processing. |
| CMCC | Proposal 1: For CSI compression using two-sided model, when using Type 1 training collaboration, the potential spec impact on AI model transfer need to be studied.  Proposal 2: For CSI compression using two-sided model, when using Type 1 training collaboration, the potential spec impact on dataset collection for training need to be studied.  Proposal 3: For CSI compression using two-sided model, when using Type 3 training collaboration, the potential spec impact on assistance signaling for AI model information need to be studied.  Proposal 4: For CSI compression using Type 1 training collaboration, whether model can be kept proprietary is up to whether the model is transferred in open format or proprietary format.  Proposal 5: For CSI compression using Type 3 training collaboration, the model could be kept proprietary.  Proposal 6: For CSI compression using Type 1 and Type 3 training collaboration, the dataset for sharing is not privacy-sensitive.  Proposal 7: For CSI compression, the model update after deployment using Type 3 training can be more flexible than using Type 1 training. |

### Summary:

For training collaboration type 3, additional dataset needs to be delivery from NW to UE in NW first training, and from UE side to NW side in UE first training. In R1-2302359, it has been proposed that at least for NW first training, the CSI generation part training dataset delivery should be over the air interface. The dataset is segmented into small piece with the same dataset ID, transmit over the air interface to many UEs. Different UEs then forward the small data piece to UE side server to re-assemble the dataset for CSI generation model training. Many companies assumed this dataset sharing can be shared between NW side and UE side transparent to RAN. Proposal 2-2-1 tries to align understanding on the necessity and methods to share the CSI generation model dataset in NW first training and/or CSI reconstruction model dataset in UE first training over the air interface.

Proposal 2-2-2 and proposal 2-2-3 are follow up proposals based on RAN1 112 agreement for NW side data collection. RAN2 is discussing data collection framework for offline training. Therefore only L1 signaling is proposed to be discussed for data collection for performance monitoring.

### ***Proposal 2-2-1:***

***In CSI compression using two-sided model use case with training collaboration type 3, for sequential training, further study necessity, feasibility, and potential specification impact on:***

* ***CSI reconstruction model training dataset and/or other information delivery from UE side to NW side for UE first training***
* ***CSI generation model training dataset and/or other information delivery from NW side to UE side for NW first training***
* ***Data sample format/type***
* ***Quantization/de-quantization related information***
* ***Note: other aspects are not precluded.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| **Google** | OK |
| CATT | Support. |
| Huawei/HiSi | Support. |
| vivo | Support |
| Ericsson | The proposal is ok with the understanding that the necessity study should be on each of these bullets individually. For example, we don’t see the necessity for using 3GPP interface of model training dataset delivery from NW to UE. |
| Xiaomi | Support |
| LG Electronics | OK |
| OPPO | Support |
| CMCC | Support. |
| Fujitsu | Support. |
| CAICT | Support. |
| Futurewei | Ok |
| Lenovo | As we have discussed in our contribution, in “iterative type 3” samples are exchanged between the UE and NW in both sides so we suggest the following change:   * ***CSI reconstruction model training dataset and/or other information delivery from UE side to NW side ~~for UE first training~~*** * ***CSI generation model training dataset and/or other information delivery from NW side to UE side ~~for NW first training~~***   Also prefer to add  “Data sample format/type **and related information, e.g., quality of the sample**” |
| NVIDIA | Support |
| ZTE | Support |
| InterDigital | Ok |
| MediaTek | Support. |

### ***Proposal 2-2-2:***

***In CSI compression using two-sided model use case, further study potential specification impact on ground truth CSI report for NW side data collection for model training:***

* ***Scalar quantization***
* ***Codebook-based quantization*** 
  + ***FFS: Parameter set enhancement of existing eType II codebook.***
* ***Whether UE or NW determine the number of ranks and the index(es) of layer(s) for ground-truth CSI report.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| **Google** | In our view, this should be deprioritized. We already have SRS to achieve the same functionality. |
| NTT DOCOMO | The third sub-bullet should be deleted. We do not find the reason why NW determines the number of ranks for ground-truth CSI instead of the maximum rank. After model training, the rank will be determined by UE based on the rank restriction. Even if NW obtains the higher rank than UE determines, that higher rank CSI will not be compressed and reported by UE. |
| CATT | Support. |
| Huawei/HiSi | Support |
| vivo | Support |
| Ericsson | Support. A key question to be answered (codebook based) and included in the proposal is whether the existing eType-II report format is sufficient or whether a more high resolution codebook for ground truth brings better trained models resulting in improved user throughput performance during inference. |
| Xiaomi | For the third bullet, it is not necessary to report the index of layers if the data of layers are reported in sequence. In addition, these three bullets are only applied to precoding matrix. Therefore, we suggest it is reworded as follows  ***In CSI compression using two-sided model use case, further study potential specification impact on ground truth CSI report for NW side data collection for model training when data type of ground truth CSI report is precoding matrix:***   * ***Scalar quantization*** * ***Codebook-based quantization***    + ***FFS: Parameter set enhancement of existing eType II codebook.*** * ***Whether UE or NW determine the number of ranks ~~and the index(es) of layer(s)~~ for ground-truth CSI report.*** |
| LG Electronics | As Google commented, we first discuss the necessity of introducing new reporting for ground-truth CSI. |
| OPPO | Support |
| CMCC | We are generally OK with this proposal.  But for the third bullet, not sure whether NW need to determine the index(s) of layer(s). Even NW determine the number of ranks *N*, UE just select the proper *N* layers, like the layers with maximum values of eigenvalues. |
| Fujitsu | Support. |
| CAICT | Support. |
| Futurewei | We are ok with Xiaomi’s update. |
| Lenovo | We suggest to add the following bullet:   * **Additional information associated with a data sample, e.g., quality of the sample**   Also, does the second bullet correspond to eType-II CSI feedback, possibly with higher-resolution/ larger parameters? |
| NVIDIA | Support |
| ZTE | Support in principle. A minor comment for clear clarification on quantization for ground-truth CSI:   * ***Scalar quantization for ground-truth CSI*** * ***Codebook-based quantization for ground-truth CSI*** |
| InterDigital | Support |
| MediaTek | Support. |

### ***Proposal 2-2-3:***

***In CSI compression using two-sided model use case, further study potential specification impact on ground truth CSI report for NW side data collection for model performance monitoring:***

* ***Codebook-based quantization*** 
  + ***FFS: Parameter set enhancement of existing eType II codebook.***
* ***L1 signaling procedure to enable fast identification of AI/ML model failure.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| **Google** | We do not think the ground truth CSI report is needed for performance monitoring. Similar to RLM/BFD, we think the performance monitoring should be based on the wireless performance, e.g., hypothetical BLER, instead of the actual CSI. From the simulation results, SGCS cannot reflect the wireless performance. |
| NTT DOCOMO | Support the proposal. |
| CATT | Support. |
| Huawei/HiSi | Support.  For the motivation of reporting intermediate KPI (ground-truth), we believe that gNB needs it to identify whether the good/bad performance is due to AI/ML model or other factors (channel status, interference, etc.), since the AI/ML algorithm may not be so robust and trustworthy as legacy UE features/algorithms.  To clarify the motivation of the “fast identification” bullet: there are two modes of monitoring: periodic monitoring and aperiodic (event-triggered) monitoring. For the latter mode, it is generally due to the sudden performance fluctuation. E.g., when the throughput of the UE running a specific UE part/UE side model suddenly degrades, gNB has to fast identify what is the reason leading to this degradation – whether it is due to the failure of the AI/ML model(s), or due to the gNB strategies of scheduling, MU pairing, etc. If it is due to the failure of the UE part/UE side model, gNB can disable it immediately. The slow L3/MDT report can hardly satisfy this requirement as its interval is up to 30min – this means the degradation persists yet gNB can do nothing but wait.  In the end, this event-triggered monitoring window expires after the identification is finished, rather than always-on, so there is no concern on overhead. |
| vivo | Support |
| Ericsson | Support. We also have the same comments is in P2-2-2. We Agree with Google that SGCS cannot reflect the performance unless for low load and rank restricted to 1 for all UEs. |
| Xiaomi | Support |
| LG Electronics | It depends on metrics and which entity performs model monitoring. Thus, necessity of ground-truth CSI report for model monitoring should be discussed first. |
| OPPO | Support |
| CMCC | We are fine with the main bullet for ground truth CSI reporting. But for quantization method, we think other quantization method is still possible, like scalar quantization. So, we suggest:   * ***Scalar quantization*** * ***Codebook-based quantization***    + ***FFS: Parameter set enhancement of existing eType II codebook.***   ***L1 signaling procedure to enable fast identification of AI/ML model failure.*** |
| Fujitsu | Support in general. We suggest that we do not restrict the signaling to L1 at this SI stage. |
| CAICT | Support. |
| Futurewei | We are ok with adding scalar quantization to the list. |
| NVIDIA | Support |
| ZTE | Generally fine with the proposal. We think ‘enable fast identification of AI/ML model failure’ in the second bullet is just an intention, which is no need to voice it and we suggest rewording it from a high level as   * ***L1 signaling procedure to ~~enable fast identification of AI/ML model failure.~~ report ground truth CSI*** |
| InterDigital | It is unclear how ground truth CSI with codebook-based quantization can help network for model performance monitoring. We tend to agree with Google that model performance monitoring should be done with other metrics. As LG proposed, we may need to discuss which metric to use for model performance monitoring and see if ground truth CSI reporting is necessary for model monitoring. |
| MediaTek | Support. |
| Intel | It is not clear for us how L1 signaling procedure to enable fast identification of AI/ML model failure related to NW-based model performance monitoring and ground truth CSI quantization. It looks like it is UE-based model performance monitoring. |

## Inference related spec impact

Following table summarize company’s proposals related to inferencing.

|  |  |
| --- | --- |
| **Company** | **Key Proposals/Observations/Positions** |
| Huawei | Proposal 11: For CQI determination of CSI compression, consider Option 1 (CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation) as a starting point.  Proposal 12: For Network configuration to determine CSI payload size, gNB can configure a set of candidate CSI payload sizes for each layer separately.  Proposal 13: For UE determination/reporting of the actual CSI payload size, UE reports the selected RI and the index of the actual CSI payload size (among the set of candidate CSI payload sizes) for each layer subject to the selected RI.   * FFS how to map the CSI report on the two parts CSI to avoid redundant feedback.   Proposal 14: In CSI compression using two-sided model, further study potential specification impact on the vector quantization and scalar quantization.   * For vector quantization,   + The format/size of the vector quantization dictionary.   + Segmentation of the encoder output.   + Configuration/reporting/updating of the quantization dictionary. * For scalar quantization,   + Uniform and non-uniform quantization.   + The configuration of the quantization granularity.   Proposal 15: For the CSI priority rules and CSI processing Unit, on top of the legacy CSI reporting principles, following AI/ML specific aspects may need to be further considered:   * The priority rules for different LCM procedures of training data collection, inference, and monitoring data collection; and the priority rule within the latent space of per CSI report. * The required CPU value of CSI calculation for per AI/ML model basis. |
| ZTE | Proposal 8: For model inference operation, further study at least the following aspects:   * Data required for model input, e.g., reference signal configurations and assistance information delivery * Report feedback based on the model output, e.g., UCI mapping order and priority * Inference latency, e.g., the relationship between inference latency and CSI reference resource   Observation 6: For CQI calculation based on target CSI with realistic channel measurement, UE may over-estimate the channel condition and reconstructed PMI and CQI are not matched. Our simulation results show that the system performance loss is obvious if no advanced CQI adjustment algorithm is used.  Observation 7: For CQI calculation based on target CSI with realistic channel measurement and adjusted by previous CSI reconstruction output provided by NW, this method needs to send back the output of CSI reconstruction part from NW side to UE, which will lead to additional latency. However, the channel condition may already change a lot (e.g., interference) so that PMI and CQI mismatch is unavoidable. In addition, the recovered CSI should be quantized (e.g., by eType II codebook), which will lead to additional quantization loss. Moreover, sending the recovered CSI needs enhanced specification to support it.  Observation 8: For CQI calculation based on target CSI with realistic channel measurement and adjusted by CQI adjustment table provided by NW, NW can construct a CQI adjustment table according to some channel characteristics based on some priori information at gNB side. Then, UE can calculate the similarity-related metrics between measured channel and the channel characteristics to do corresponding CQI adjustment.  Observation 9: For CQI calculation based on legacy codebook, UE my not support traditional codebook and AI/ML codebook simultaneously, which will largely increase the UE complexity. Meanwhile, PMI and CQI mismatching is also unavoidable. If traditional codebook can already get accurate PMI, it is not necessary to implement AI/ML models.  Observation 10: For CQI calculation based on CSI reconstruction output, where CSI reconstruction part at the UE is the same as the actual CSI reconstruction part used at the NW. UE may also be not expected to have CSI reconstruction model as it increases UE computation/storage/power consumption burden to a large extent. In addition, the CSI reconstruction model is generally a proprietary design by network side.  Observation 11: For CQI calculation based on the output of CSI reconstruction model assumed at UE, this method can be applicable for CSI compression using two-sided model sub use case and shows that the average system UPT can be achieved almost the same as the case of CQI calculation based on the output of actual CSI reconstruction model (i.e., performance upper-bound for all options).  Observation 12: For CQI calculation using two stage approach, it is already supported (i.e., when the report quantity is cri-RI-CQI) and less specification impact is foreseen. Besides, the two-step procedure increases the time span of the CQI determination process, which may face the channel variation/aging so that the current CQI cannot match the previous CSI.  Proposal 9: The performance of different CQI determination options should be evaluated in agenda item 9.2.2.1.  Proposal 10: Further categorize the Option 1b as following:   * Option 1b-1: CQI is calculated based on target CSI with realistic channel measurement and adjusted by previous CSI reconstruction output provided by gNB * Option 1b-2: CQI is calculated based on target CSI with realistic channel measurement and adjusted by CQI adjustment table provided by gNB.   Proposal 11: Further categorize the Option 2a as following:   * Option 2a-1: CQI is calculated based on CSI reconstruction output, where CSI reconstruction part at the UE is the same as the actual CSI reconstruction part used at the NW. * Option 2a-2: CQI is calculated based on CSI reconstruction output assumed at UE side, where CSI reconstruction part at the UE is different from the actual CSI reconstruction part used at the NW.   Proposal 12: According to initial evaluations on performance and specification impacts, the following down-selections are proposed:   * Further study the specification impacts (including the feasibility and necessity) on Option 1a, Option 1b-2 and Option 2a-2. * No further discussion on specification impacts for Option1b-1, Option 1c, Option 2a-1 and Option 2b.   Proposal 13: In CSI compression using two-sided model use case, LI determination should be studied along with CQI determination.  Proposal 14: In CSI compression using two-sided model use case, if RI is configured to be reported, legacy RI determination can be reused as a starting point.  Proposal 15: Further study potential specification impact on more channel information reported for MU-MIMO scheduling, e.g., full rank report based on the AI/ML model.  Proposal 16: In CSI compression using two-sided model use case, further study the following quantization alignment options:   * For scalar quantization scheme, the quantization dictionary should be aligned including quantization type, quantization level, quantization pattern, etc. * For vector quantization scheme, the quantization codebook should be aligned including the length of codeword, the size of codebook, etc. * The configuration/reporting/update of the quantization dictionary/codebook.   Proposal 17: The performance of different monitoring cases based on intermediate KPIs and the related evaluation KPIs should be discussed for companies to compare the monitoring performance in agenda item 9.2.2.1.  Proposal 18: Further study the specification impacts on least the following two cases for model performance monitoring,   * Case 1: UE-side monitoring based on the output of the CSI reconstruction model assumed at the UE-sid , e.g., Intermediate KPIs are calculated by UE based on the output of the CSI reconstruction model assumed at the UE side, which is not the same as the actual CSI reconstruction model used at the NW side. * Case 2: NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE or obtained from the UE-side, e.g., Intermediate KPIs are calculated by NW based on traditional CSI and CSI reconstruction model output.   Proposal 19: For NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, further study a high-resolution CSI based on traditional codebook as ground-truth label.  Observation 13: For training type 3, CSI generation model and CSI reconstruction model are actually two separate models. Therefore, if the performance of output CSI is degraded, it cannot be decided whether it’s due to the performance loss of CSI generation model or CSI reconstruction model.  Proposal 20: Further study the potential mechanisms and specification impacts on monitoring model performance of the CSI generation model and CSI reconstruction model separately.  Proposal 21: Deprioritize the model performance monitoring based on eventual KPIs.  Proposal 22: Further study the feasibility of input/output-based monitoring methods in Agenda item 9.2.2.1.  Proposal 24: In CSI compression using two-sided model use case, further study the methods and potential specification impact on mapping priority and omission rule for AI/ML CSI report,   * Dynamic quantization resolution to reduce payload * Divide the CSI into multiple groups with different priority and omit the CSI groups with low priority, e.g., according to layer, subband, port * CSI reporting is separated into multiple reports, e.g., to establish the association among the multiple reports |
| OPPO | * Proposal 2: CQI should be calculated based on target CSI with realistic channel measurement. * Proposal 3: Regarding the CSI input,   + when UE obtains the CSI generation part from NW in a 3GPP non-transparent way, the network needs to explicitly or implicitly indicate the input interface format of the CSI generation part, e.g. data type, dimension size, normalization/quantification schemes.   + when UE obtains the CSI generation part in a 3GPP transparent way, no need to indicate the input interface through 3GPP protocols * Proposal 4: The training complexity, inference complexity, signaling cost for indication and standardization impact of different quantization/dequantization methods need to be evaluated.   + If the quantization/dequantization scheme is not a key contributor to CSI compression/recovery performance, the quantization/dequantization scheme(s) that is relatively simple, easy to indicate and have less standardization impact(e.g. case 2-1) should be selected first. * Proposal 5: In CSI compression using two-sided model use case, further study potential specification impact on the quantization/dequantization method for the compressed CSI, including   + At least for training collaboration type3, quantization/dequantization methods should be specified and aligned to ensure the encoder and encoder to be well trained and could work together     - For NW first training, network should indicate the quantization [or the dequantization] method for the compressed CSI to UE.     - For UE first training, UE should indicate the dequantization [or the quantization] method for the compressed CSI to NW.   + Study potential signaling and procedure to indicate the quantization/dequantization method |
| vivo | 1. Quantization-non-aware training for CSI compression would suffer from a significant performance loss compared with Quantization-aware training. 2. If quantization method at CSI generation part and dequantization method at CSI reconstruction part are not aligned, there will be an unacceptable performance loss for AI/ML models. 3. Study the potential specification impact of the alignment of quantization method at UE side and dequantization method at NW side based on different training collaboration types for CSI compression. 4. Similarity and orthogonality loss can be used for CQI adjustment based on target CSI with realistic channel measurement 5. It is possible for AI/ML models in CSI compression to support the priority rule regarding CSI collision handling and CSI omission if payload truncation is considered during training. 6. Study the feasibility and specification impacts for AI/ML models in CSI compression to support the priority rule regarding CSI collision handling and CSI omission. Considering payload truncation during training can be set as one starting point. 7. Legacy codebook subset restriction (CBSR) framework can be directly supported in AI/ML based CSI compression by constraining the input CSI towards particular direction while reusing the same model as ordinary cases. 8. Study the CSI processing Unit design for AI/ML based CSI compression. |
| Spreadtrum | Proposal 9: Aperiodic CSI reporting should be considered firstly.  Proposal 10: The configuration of CSI-ResourceConfig and/or CSI-ReportConfig should be enhanced  Proposal 11: CQI/RI still should be included in the CSI report.  Proposal 12: Regarding CQI calculation, option 1a and/or option 1b can be considered.  Proposal 13: The priority for AI/ML based CSI feedback needs to be considered.  Proposal 14: Introducing for CSI reports carrying CSI compression information enabled by AI/ML operation in the priority rule for CSI reports.  Observation 2: Codebook subset restriction can be not considered in CSI compression and recovery using two-sided model use case.  Proposal 15: How to define/reflect the complexity of the AI/ML operation in the specification should be considered. |
| Nokia | Proposal 5: Regarding the quantization scheme for CSI feedback, a scalar quantization scheme with a limited bit size needs to be studied especially for bounded input to the AI encoder use case, e.g., channel eigenvector compression.  Proposal 6: Regarding vector quantization scheme for CSI feedback for Type 2 or Type 3 two-sided model training collaboration scenarios, the degree of required alignment between quantizer/dequantizer at UE-side/NW-side respectively needs to be studied, e.g., the length of a codeword, the size of a codebook, and the distance metric (or quantization rule) in use.  Observation 1: The size of VQ codebook can cause limitations/difficulties in using VQ and needs to be investigated.  Proposal 7: RAN1 may investigate sharing the relevant quantization architecture and parameters from one network entity to the other. For example, the type of quantization and quantization parameters can be shared with the other network node. The quantization parameters depend on the quantization type and may include:   * For scalar uniform quantization: number of quantization bits/levels, the minimum and maximum range of quantization * For scalar non-uniform quantization: number of quantization bits/levels, the minimum and maximum range of quantization, type of non-linear function and its parameters * For vector quantization: Codebook size and all the codewords   Proposal 10: RAN1 shall study the possible specification changes when supporting multiple compression ratios and how to enable progressive training.    Proposal 11: RAN1 shall study the possible specification changes when accommodating various CSI-RS configurations (e.g., bandwidths, ports) and multiple payload sizes.  Proposal 12: RAN1 shall study the possible use of CSI part 1 and CSI part 2 like approach for the compressed CSI reporting. |
| CATT | Proposal 14: For AI/ML based CSI feedback, the overheads of CSI feedback for rank 3 and rank 4 are expected to be comparable with that of rank 2.  Proposal 15: In CSI compression using two-sided model use case, if CQI in CSI report is configured, for CQI determination in CSI report, one of the sub options of Option 1 is adopted:   * Option 1: CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation, including   + Option 1a: CQI is calculated based on target CSI with realistic channel measurement   + Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment   + Option 1c: CQI is calculated based on legacy codebook   Proposal 16: For CQI reporting in CSI compression using two-sided model use case, the same quantization schemes as that in Rel-17 for codebook based CSI feedback is considered. |
| Intel | Proposal 5:   * It is expected that AI/ML model is trained assuming a particular pre/post processing   + If an AI/ML model is configured at the UE for inference, information on pre-processing for that model should be provided to the UE (e.g. specified, configured, downloaded etc.)   + Pre/post-processing may include at least linear transforms (DFT across different dimensions), down selection of matrix elements and normalization   Proposal 6:   * The dimensions of the input are defined by parameters similar to parameters L/M for Enhanced Type II PMI codebook (considering that input corresponds to the neural network input after pre-processing)   + In some cases, information from pre-processing step shall be reported by the UE together with CSI bits generated by the neural network (e.g., selected basis vectors, basis rotation factor, etc.)   Proposal 7:   * Consider existing principles for RI and CQI for spatial-frequency domain CSI compression using two-sided AI model sub-use case   Proposal 8:   * The following alternatives for CQI adjustment determination can be considered for Option 1b CQI determination   + CQI adjustment is configured via higher layers   + CQI adjustment is determined by the UE based on reference CQI (e.g., measured from precoder CSI-RS)   + CQI is calculated using precoding matrix corresponding to the target CSI with added AWGN |
| Interdigital | Proposal 1: Perform a trade-off analysis of the performance, complexity and standardization impacts of both precoding matrix and explicit channel matrix before prioritizing.  Observation 1: Different pre-processing types are beneficial under different deployment scenarios and channel characteristics.  Observation2: Different pre-processing types lead to different AI/ML encoder outputs which need to be known at the decoder.  Proposal 2: Study support of multiple pre-processing options.  Proposal 3: Study UE selection and reporting of pre-processor type.  Proposal 4: Study UE determination and reporting of the RI and CQI based on the input to the AI/ML model at the UE.  Observation 3: A UE without an up-to-date AI/ML decoder cannot independently detect CQI mismatch.  Proposal 5: Study means to detect and identify when there is mismatch between a UE’s AI/ML encoder input and the NW’s AI/ML decoder output.  Proposal 6: Study methods to enable CQI adjustment based on detected CQI mismatch.  Proposal 7: Study specification impacts of CSI compression using AI/ML including: CSI compression type, support of multiple AI/ML models, new CSI reporting mechanisms and fallback to legacy CSI reporting. |
| Interdigital | Proposal 16: Study quantizer/dequantizer updating separate from AI/ML model switching.  Proposal 17: Study different alignment levels between quantizer and dequantizer.  Proposal 18: For models with quantization non-aware training, study non-uniform quantization as means to determine actual CSI payload size within the NW configured constraints. |
| Fujitsu | Proposal-1: For the sub use case of CSI compression using two-sided AI/ML models, study the mechanism that UE and NW align their supported AI/ML models in the multi-vendor collaborations.  Proposal-2: For the sub use case of CSI compression using two-sided AI/ML models, study the method for indicating the pairing information of the AI/ML-based CSI generation and reconstruction parts. The pairing ID can be studied as a starting point. |
| Ericsson | [Proposal 7 Target CSI is standardized by use of the implicit CSI reporting principle (precoding vector) and is based on the eType-II framework. Study further the parameter values, e.g., of L, p\_v, β,..](#_Toc131752944)  [Proposal 8 RAN1 to study whether the number of quantization levels per encoder output should be fixed or configurable by the network in CSI report configuration.](#_Toc131752945)  [Proposal 9 Re-use the legacy CSI reporting principle with CSI Part 1 and Part 2 where Part 1 has a network configured fixed size and Part 2 size is dynamic, determined by information in Part 1.](#_Toc131752946)  [Proposal 10 The UCI for an AI-CSI report consists of bits carried in CSI part 1 for the auxiliary information common across all the transmission layers, bits carried in CSI part 2 used to complete the interpretation of the output CSI, and bits carried in CSI part 2, representing the quantized latent space output of the encoder.](#_Toc131752947)  [Proposal 11 Support Option 1 with CQI being calculated based on a hypothetical CSI which is derived from target CSI. Further study the details of mechanisms for CQI adjustments.](#_Toc131752948)  [Proposal 12 If target CSI being an explicit channel tensor is supported (i.e. full Tx \* Rx MIMO channel), an alternative solution is that the CSI report doesn’t contain CQI and RI, but contains an interference plus noise (IpN) report.](#_Toc131752949)  Observation 9: Given the potential complexity arising from unmatched quantization, proponents of non-standardized quantization need to motivate the benefits to why the quantization should not be standardized.  Observation 10 : It is feasible to have a quantization-common model, the performance difference to a quantization-specific model is non-substantial.  Observation 11: If the pre-processing contains removal of raw channel subspace (by the UE), then information about the remaining subspace (e.g., the SD and FD basis vectors) needs to be reported to the network side along with the encoder output bits.  Observation 12: The importance of CBSR will increase due to more complicated interference situations in coming deployments and bands  Observation 13 : A benefit of a Target CSI definition based on eType-II is that CBSR can straightforwardly be applied by gNB to UE configuration of the target  Observation 14 : Since a CSI-RS measurement may be used for multiple purposes (monitoring, inference, data collection), and processed by different hardware in the UE, RAN1 can consider discussing CPU and measurement processing units (MPU) as two decoupled entities used to define the UE processing load |
| xiaomi | Proposal 5: Alt2b, i.e., CQI is calculated using two stage approach, where UE derives CQI using precoded CSI-RS transmitted with a reconstructed precoder should be supported.  Proposal 6: The legacy priority rule can be reused to define the priority the AI/ML based CSI reporting, and a priority value with new parameter value or introducing new parameter is used to indicate the priority of CSI reporting.  Proposal 7: CSI reporting with two parts, i.e., Part 1 and Part 2 or only one part for AI/ML based CSI feedback with two-sided model can be supported.  Proposal 8: The compressed quantization information is divided into N>1 groups for CSI omission, where the values N and how to divide into N groups needs to further study.  Proposal 9: RAN 1 should study the AI processing unit capability report and AI processing unit number determination for various cases of AI based CSI enhancement. |
| Panasonic | Observation 15: For each option of training collaboration, handling of rank of AI/ML model should studied.  Observation 16: Both quantization non-aware training and quantization-aware training should be studied.  Observation 17: For CQI determination in CSI report, further study Option 1a, 1b, and 2a.  Observation 18: Legacy CSI reporting mechanism, i.e., mapping of compressed CSI into fixed/configurable/known-payload part (similar to CSI part 1) and variable/predictable size (similar to CSI part 2) may also be required for CSI compression using two-sided models. |
| LGE | Proposal #3: For CSI reporting for AI/ML based CSI compression, two-part encoding can be considered where # of actual bits for AI/ML generated CSI can be included in Part 1 CSI. FFS on it can be across layer or per layer.  Proposal #4: For CQI determination of AI/ML based CSI compression, prioritize option 1 (CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation).  Proposal #5: Consider CSI compression ratio information as CSI reporting contents.  Proposal #6: Consider enhancement of CSI restriction at least followings   * Configuration associated with form of ids such as configuration id, site id, zone id, etc. * Dynamic configuration switching   Proposal #7: Consider defining new CSI processing unit to handle the AI/ML based CSI. |
| ETRI | Proposal 2: In CSI compression using two-sided AI model, further study the following potential specification impacts on UCI configuration.   * NW configures UE to generate the UCI payload in a certain size. * UE generates the UCI payload within the maximum UCI payload size. UE delivers to or shares details of the UCI payload (including quantization-related information)   Observation 3: By setting asymmetric quantization levels for the encoder output allows dynamic adjustments of UCI payload. |
| CMCC | Proposal 8: In AI based CSI compression, for NW side data collection, the following two high resolution quantization methods could be considered for ground-truth CSI reporting:  • High resolution scalar quantization, e.g., Float32, Float16, etc.  • High resolution codebook quantization, e.g., R16 Type II-like method with new parameters  Proposal 9: For CSI compression using two-sided model use case, the enhancement on CSI processing time and the definitions of Z and Z’ could be studied. |
| MediaTek | 1. For VQ, UE and gNB should align their codebook and segmentation approach. 2. Prioritize SQ methods over VQ. |
| Apple | Proposal 4: For eigen-vector based CSI compression, NW configure the maximum UCI size and list of candidates NN IDs via RRC configuration.  Proposal 5: For eigen-vector based CSI compression, the UE determine which AI model to use based on rank and include the model ID as part of the CSI report.  Proposal 6: To enable quantization alignment in AI based CSI compression training type 3, specify at least the size/dimension of CSI generation model output before quantization.  Proposal 7: If vector quantization is supported in AI based CSI feedback, the input/output size of vector quantization codebook should be specified.  Proposal 8: When domain transformation pre-processing is used, legacy CSI report principle can be applied to input CSI directly.  Proposal 9: When domain transformation pre-processing is not used,   * Prioritization rule is indirectly support by selecting different AI model with different UCI bit size. * CBSD can be supported by projecting the input CSI in the subspace orthogonal to restricted sub-space before AI model. |
| Lenovo | 1. The quantization/dequantization method of the AI/ML model output is pre-configured prior to CSI feedback process 2. Study different alternatives for quantization/dequantization methods for CSI compression, considering rank common/specific design, as well as layer common/specific design 3. Study different alternatives of reporting the AI-based CSI framework configuration parameters based on the design details of the AI-based CSI compression framework 4. Study potential CSI report characteristics for AI-based CSI compression under different network-UE training collaboration levels 5. Prioritize Option 1a and Option 2a for CSI compression format in spatial-frequency domain 6. For the mapping order of CSI fields corresponding to AI-based spatial-frequency CSI compression, the CSI feedback is composed into two parts:  * Part 1: comprising RI, CQI and size of CSI Part 2, where the size of CSI Part 1 is fixed * Part 2: comprising the AI encoder output, where the size of Part 2 is indicated in CSI Part 1  1. Strive to design the AI-based spatial-frequency CSI compression codebook so that (i) the overall CSI feedback is fixed for different *RI* values and/or different channel conditions, or (ii) the CSI fields are mapped in an order that enables partial UCI omission of the CSI feedback without jeopardizing the un-omitted CSI feedback 2. Assuming two-sided AI models for CSI compression under training collaboration Type 3, further enhancements are needed to ensure precise CQI characterization in the presence of mismatch between the nominal decoder at the UE side and the actual decoder at the network side 3. Consider Option 1b for CQI reporting, where the UE appends side information to the CQI calculated based on the nominal decoder, such that the side information helps quantify the encoder/decoder mismatch to enable more accurate CQI adjustment to the actual CQI value 4. CBSR is supported for AI-based CSI reporting 5. Reuse legacy DFT-based CBSR, where a DFT-based restricted vector *r* implies that no precoding vector *v* within a pre-determined angular distance from the vector *r* can be selected |
| Qualcomm | Observation 16: Only UCI and final format of the reported CSI (e.g., the precoding matrix) are specified in legacy CSI feedback framework. The PMI search algorithm and its input are proprietary.  Observation 17: In CSI feedback via two-sided model, PMI searching algorithm is replaced by UE-side model while PMI codebook is replaced by NW-side model. The general principle for specification impact should be preserved. The need for specifying UE-side input and pre-processing is not clear.  Observation 18: Post-processing of NW-side model output into the final CSI format can be absorbed into the specification of the final CSI format.  Observation 19: Channel matrix feedback (i.e., H-in-H-out) creates additional and unnecessary complexity for multi-vendor operation.  Observation 20: Eigen-value or soft-rank feedback, along with precoder, achieves similar merit as the channel matrix feedback in terms of flexibility for network scheduling without causing significant increase in implementation complexity.  Observation 21: Quantization non-aware training (case-1) leads to noticeable performance degradation compared with quantization aware training (case-2).  Observation 22: Trainable quantization offers more flexibility and better performance compared to fixed quantization, e.g., trainable vector quantization can improve the performance.  Proposal 15: Reuse current CSI report configuration framework with new signaling of pairing ID and necessary information related to the CSI feedback, e.g., rank restriction, antenna port configuration, payload information.  Note: A pairing ID is a logical ID that indicates compatibility between the UE-side and NW-side model of a two-sided model. For example, all encoders developed from a two-sided multi-vendor training session may be associated with a single pairing ID. As another example, in NW-side first training, UE-side encoders trained based on the same NW-side model may be associated with a single pairing ID.  Proposal 16: Study payload scalability with number of subbands, number of ports and rank.  Proposal 17: UE-side actual payload determination should be based on only reported rank for two-sided ML-CSI feedback.  Proposal 18: The input to the UE-side model should be left to UE implementation, the output at the NW-side model can be specified.  Proposal 19: Preprocessing at UE-side is upto UE-implementation and should not be specified.  Proposal 20: For AI-based CSI feedback, the size of the UCI payload and the final CSI format can be specified.  Proposal 21: Study reporting the precoding matrix together with eigen-values or soft-rank for two-sided AI/ML CSI feedback.  Proposal 22: Deprioritize channel matrix feedback for the R18 study item.  Proposal 23: Quantization method should be considered a part of the UE-side model and dequantization method should be considered a part of the NW-side model. The quantization method should be aligned for good performance, but there is no need for separate specification support to align the quantization method. |
| NTT DOCOMO | Observation 5: Model and assistance information can be used for paring of trained two-sided models in CSI compression.  Observation 6: There is another mechanism to help MCS selection, such as HARQ-ACK mechanism, in addition to CQI reporting.  Observation 7: For CSI compression, the constraint on channel for CSI reporting can be the same as subband type II codebook.  Observation 8: For CSI compression, CSI reporting can consist of two parts; CSI part 1 including RI/encoder model ID/CQI, and CSI part 2 including compressed bits.  Observation 9: For CSI compression, the legacy priority rules for CSI collision handling and CSI omission can be reused except for the priority reporting level within the compressed bits.  Observation 10: NW side monitoring with target CSI reporting suffers from the signalling overhead and quantization error of target CSI reporting.  Observation 11: UE side monitoring with NW indication of reconstructed CSI suffers from the signalling overhead and quantization error of reconstructed CSI indication.  Observation 12: The feasibility of UE side monitoring with reconstruction model at UE side is questionable due to the additional model storage and processing for the reconstruction model at UE side.  Observation 13: Empirical system performance does not require the additional signalling and measurement. However, the relevance to the model performance is low compared to the inference accuracy KPI.  Proposal 3: Clarify what model is identified by model ID in the two-sided model. Until the clarification is made, it is better to introduce paired model ID, encoder model ID, and decoder model ID for the discussion purpose.  Proposal 4: Study the mechanism to align the paired trained models for two-sided models.  Proposal 5: Reuse legacy CSI reporting principle, unless technical issue is observed.  Proposal 6: It is unnecessary to explicitly indicate/configure the CSI payload size. Instead, CSI payload size can be implicitly calculated based on the rank and model ID/functionality information.  Proposal 7: CQI calculated based on target CSI is sufficient for CSI compression.  Proposal 8: For CSI compression, the legacy priority rules for CSI collision handling and CSI omission can be reused except for the priority reporting level within the compressed bits. |
| Samsung | Proposal 2-2: For AI/ML based CSI compression sub-use case, study the specification impact of UCI format for quantized output of CSI generation part.  Proposal 2-3: For AI/ML based CSI compression sub-use case, study flexible configuration of quantization method and quantization resolution that enables the network to  1) Adapt to different AI/ML models and channel environments/scenarios  2) Control the feedback payload size.  Proposal 2-4: For AI/ML based CSI compression sub-use case, study the specification impact of adaptable CSI feedback payload size that enables the UE to adapt to available size of uplink resources.  FFS: whether priority and CSI dropping rules have to be introduced.  Proposal 2-5: For AI/ML based CSI compression sub-use case, study methods to configure and apply codebook subset restriction (CBSR) including:   * Whether the legacy SD basis vectors based restriction applies * How to apply CBSR for when Output-CSI-UE is in 1) spatial-frequency domain 2) angle-delay domain * Whether soft amplitude restriction is possible   Proposal 2-10: In CSI compression using two-sided model, adopt Option 1a: CQI is calculated based on target CSI with realistic channel measurement.  Observation#1  In case of MU-MIMO, the network may not directly apply the precoder based on reported PMI, e.g., for interference nulling, etc. Thus, even in legacy systems, some level of mismatch exists between the PMI (precoder network reconstructs from PMI) and the precoder network applies for data transmission.  Observation#2  In CSI compression using two-sided model, for CQI determination in CSI report, for Option 1a: CQI is calculated based on target CSI with realistic channel measurement   * + Is computationally friendly as UE does not require to perform CSI reconstruction or additional measurements for CQI calculation   + The mismatch between CQI determined conditioned on target CSI (precoder) and CQI determined conditioned on the reconstructed CSI (precoder) is insignificant when CSI reconstruction loss is insignificant   Observation#3  In CSI compression using two-sided model, for CQI determination in CSI report, for Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment:   * + The adjustment can be handled in a spec. transparent manner.   Observation#4  In CSI compression using two-sided model, for CQI determination in CSI report, for Option 2a: CQI is calculated based on CSI reconstruction output, if CSI reconstruction model is available at the UE and UE can perform reconstruction model inference with potential adjustment:   * + The availability of Network’s reconstruction output at the UE is not guaranteed, as network may be willing to share it, thus, may not be feasible.   + Network may use heavier model, which may not fit in to UE’s computational capability, thus, may not be feasible.   Observation#5  In CSI compression using two-sided model, for CQI determination in CSI report, for Option 2b: CQI is calculated using two stage approach in which UE derives CQI using precoded CSI-RS transmitted with a reconstructed precoder:   * + It incurs additional CSI-RS overhead   + The delay between CSI (precoder) generation and CQI determination introduces mismatch. |

### Summary:

Quantization is an essential part of the CSI compression use case. Alignment of quantization methods are required. It is important to identify the aspects that required standardization to facilitate multi-vendor joint training procedure. Proposal 2-3-1 summarize companies’ proposal on potential spec impact related to quantization.

In legacy CSI feedback framework, the gNB has the flexibility to configure the maximum CSI feedback payload size by configuring the maximum rank number, the codebook type, CSI feedback granularity and the codebook parameter combinations. The UE can autonomously determine the RI and the number of non-zero coefficients which are fed back to the gNB so that the UE also has the flexibility of determining the CSI feedback payload and adapting the UCI report based on radio environment. For AI/ML-based solutions, methods to enable similar level of flexibility of configuring/determining the CSI payload size by both gNB and UE need further discussion. Related information to ensure paired AI models are used is one main aspects of CSI configuration and reporting.

The logic model ID linking paired CSI generation model and CSI reconstruction model need to be defined for CSI configuration and reporting purpose. For training collaboration type 3, with 1 to M training, it is possible that one decoder maps to multiple encoders. It is also possible that different encoders are trained for different decoders in NW first training. In addition, based on the assisted information in categorizing the dataset, different encoder and/or decoder are trained for different dataset or common encoder/decoder are trained for different dataset. The logical model ID needs to flexibly and efficiently indicate encoder/decoder pair to ensure correct inferencing (reference R1-2303706).

Proposal 2-3-2 and 2-3-3 discuss the CSI configuration and model ID options.

### ***Proposal 2-3-1:***

***In CSI compression using two-sided model use case, further study at least the following potential specification impact on quantization alignment including:***

* ***For vector quantization scheme,*** 
  + ***The format and size of the VQ codebook, the distance metric (or quantization rule)***
  + ***Size and segmentation method of the CSI generation model output***
  + ***Configuration/reporting/updating of the VQ codebook***
* ***For scaler quantization scheme,***
  + ***Uniform and non-uniform quantization***
  + ***Configuration of the quantization granularity.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | We think the quantization should be part of model, which does not require spec impact. |
| NTT DOCOMO | Same view as Google. |
| CATT | It seems the proposal depends on how the model is trained, i.e. quantization -aware or non-quantization aware. So quantization is possible to be part of the model. |
| Huawei/HiSi | Support the proposal in principle, except for the distance metric: it should be implementation at the training entity or the VQ CB updating entity? As long as the VQ CB is finished, the updating entity only needs to deliver the CB to the other side. |
| vivo | Support |
| Ericsson | Support |
| Xiaomi | According to our simulation, there are no much difference for the two quantization methods. Therefore, we can firstly discuss which one is adopted based on the evaluation results in AI9.2.2.1. |
| LG Electronics | Fine with the proposal. |
| OPPO | Whether it is necessary to consider vector quantization methods depends on the evaluation results in 9.2.2.1.  If the quantization/dequantization scheme is not a key contributor to CSI compression/recovery performance, the quantization/dequantization scheme(s) that is relatively simple, easy to indicate and have less standardization impact(e.g. scaler quantization scheme) should be studied with high priority. |
| CMCC | We think the quantization part and the dequantization part should be paired well so that the whole performance can be good. At least for quantization -aware training, the quantization part should be part of model and there seems no extra spec impact. |
| CAICT | Fine with the proposal. |
| Futurewei | We agree with Huawei that the training/updating entity will train/update the VQ codebook then transfer/deliver to the other entity; distance function doesn’t need to be explicitly share between entities, i.e., gNB and UE. |
| NVIDIA | Support |
| ZTE | We are fine with the proposal.  For VQ codebook, to our understanding, the distance metric may be an implementation manner, which may not need specification. We suggest removing it as   * ***For vector quantization scheme,***    + ***The format and size of the VQ codebook, ~~the distance metric (or quantization rule)~~***   In addition, a minor typo should be revised:   * ***For ~~scaler~~ scalar quantization scheme,***   + ***Uniform and non-uniform quantization***   + ***Configuration of the quantization granularity.*** |
| InterDigital | OK with proposal |
| MediaTek | Support. |

### ***Proposal 2-3-2:***

***In CSI compression using two-sided model use case, further study the potential specification impact for CSI configuration and report:***

* ***For Network configuration to determine CSI payload size, gNB can configure a list of model ID indicating the potential CSI generation models UE can choose.*** 
  + ***FFS: whether the configuration is per layer or common to all layers***
  + ***FFS: the model ID format***
* ***For UE determination/reporting of the actual CSI payload size, UE reports the selected RI and the model ID indicating the corresponding CSI reconstruction model for each layer subject to the selected RI.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | We think there is similar problem for RI report as CQI report. We need to firstly study whether to support one-stage or two-stage report for RI. CQI can be compensated by OLLA, but RI may not. Thus, two-stage report could be more beneficial for RI report. We suggest the last bullet as follows:   * ***For UE determination/reporting of the actual CSI payload size, study the following options:***   + ***Option 1: UE reports the selected RI and the model ID indicating the corresponding CSI reconstruction model for each layer subject to the selected RI.***   + ***Option 2: UE reports the model ID indicating the corresponding CSI reconstruction model for each layer subject to the maximum number of DL layers.***     - ***The gNB may configure another CSI report including at least the rank related report*** |
| NTT DOCOMO | Support the proposal |
| CATT | While we think further study is OK, the wording seems only considers model-ID based LCM.  If the group would like to support two-sided model in functionality-based LCM, ‘model ID’ should better be rewritten in a more general way, e.g. pairing ID used in agenda 9.2.1:  ***In CSI compression using two-sided model use case, further study the potential specification impact for CSI configuration and report:***   * ***For Network configuration to determine CSI payload size, gNB can configure a list of ~~model~~ pairing ID indicating the potential CSI generation models UE can choose.***    + ***FFS: whether the configuration is per layer or common to all layers***   + ***FFS: the ~~model~~ pairing ID format*** * ***For UE determination/reporting of the actual CSI payload size, UE reports the selected RI and the ~~model~~ pairing ID indicating the corresponding CSI reconstruction model for each layer subject to the selected RI.***   ***Note: model ID may serve as pairing ID in model-ID based LCM.*** |
| Huawei/HiSi | We think the proposal in this form only address one of the possibilities.  Technically, there can be a number of solutions to support various CSI payload sizes: 1) developing scalable models, 2) configure various granularities of the SQ/VQ CB, etc., while 3) configuring a number of models is one of them.  Model-based LCM is still under discussion in 9.2.1. At this stage, to make a unified solution regarding the determination of the CSI payload size, we think this issue should be decoupled from the model ID discussion. gNB can simply configure a set of candidate CSI payload sizes and UE reports the index of the actual CSI payload size along with the selected RI. |
| vivo | This proposal implies that the model with different CSI payload size needs different model ID for CSI generation models. One way in our discussion, however, is that UE can use a single CSI generation model by means of payload truncation mechanism to fit the payload, and NW needs a CSI deconstruction model associated with the truncated payload. This works with a single model ID, as long as UE implicitly/explicitly informs the payload size of reporting to NW. |
| Ericsson | Agree with CATT that pairing ID should be used to identify a paired set of models (plus used dataset) from a training collaboration process.  We Cannot support the proposal to signal a list, as the gNB only use one decoder, i.e. one pairing ID at a time due to complexity reasons with handling multiple parallel models. So there is no need to signal a list to the UE since the UE cannot freely switch between different encoder models. Agree with Huawei there are many different options to adapt the overhead. |
| Xiaomi | Before discussing the proposal, we should have the common understanding whether the two-sided model is managed by functionality identification or model identification. |
| LG Electronics | We are not sure on the necessity of configuration of a list of model IDs for CSI payload determination and reporting of model ID. Similar to CQI determination, RI also can be calculated based on the realistic channel estimation. |
| OPPO | We prefer keep this discussion open, decoupled from the model ID discussion and collect more views and candidate solutions. |
| CMCC | Agree with Huawei. There might more than one option to deal with multi ranks issues. One model could be scaled with variable ranks. |
| Fujitsu | Support. |
| CAICT | Prefer to discuss the details of CSI configuration and report with more conclusions from 9.2.1 on Model ID. |
| Futurewei | Whether/how functionality-based LCM and Model ID based LCM will be used is still being discussed in 9.2.1, so it’s better to use general term (e.g., a list of CSI payload sizes) vs. “Model ID”. |
| Lenovo | The network configuration bullet is not clear. Does the UE determine the RI? If so, the CSI payload should be roughly the same for all RI values. If the rank is determined by the network, further justification is needed on how the network can precisely compute the best rank, at least for FDD where reciprocity does not hold |
| NVIDIA | This proposal appears to be too early to discuss. |
| ZTE | Agree with Huawei that there are multiple options for achieving different CSI payload size. In addition, for Network configuration to determine CSI payload size, gNB can also configure the specific model for UE to apply, not necessarily a list of models. For ‘model ID’, we think it should be replaced with ‘**pairing ID**’ since functionality-based LCM should also be considered. |
| InterDigital | Agree with some other companies that it would better to decouple it from the model ID if we want to progress further. Or, wait until we have solid understanding on model identification and functionality identification. |
| MediaTek | Support. |

### ***Proposal 2-3-3:***

***In CSI compression using two-sided model use case, further study the following options to define the model ID for CSI configuration and report:***

* ***Option 1: The model ID indicates the CSI reconstruction model ID that NW will use.***
* ***Option 2: The model ID indicates the CSI generation model ID that the UE will use.***
* ***Option 3: The model ID indicates the paired CSI generation model and CSI reconstruction model.***
* ***Option 4: The model ID indicates by the dataset ID during training type 3 offline training.***
* ***Other options are not excluded.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | We think we can add “reference” between the word “model” for each option. The UE and NW should communicate based on some logical model instead of actual/physical model. |
| NTT DOCOMO | We think option 2 is reasonable and consistent with one-sided model. The decoder paired to each model ID can be included in mete info. That way, NW can implicitly indicate the NW side model via indicating the model ID that represents the encoder. |
| CATT | Generally OK.  As we suggest in previous proposal, ‘model ID’ is better to be rewritten as ‘pairing ID’ to incorporate functionality-based LCM. This is also makes Option 4 less confusing…  ***In CSI compression using two-sided model use case, further study the following options to define the ~~model~~ pairing ID for CSI configuration and report:***   * ***Option 1: The ~~model~~ pairing ID indicates the CSI reconstruction model ID that NW will use.*** * ***Option 2: The ~~model~~ pairing ID indicates the CSI generation model ID that the UE will use.*** * ***Option 3: The ~~model~~ pairing ID indicates the paired CSI generation model and CSI reconstruction model.*** * ***Option 4: The ~~model~~ pairing ID indicates by the dataset ID during training type 3 offline training.*** * ***Other options are not excluded.*** |
| Huawei/HiSi | We think this proposal is tangled with the model/functionality identification, and should be discussed in 9.2.1 along with other issues related to the model-based LCM.  E.g., in our understanding, the dataset ID in Option 4 can operate under the functionality identification manner. |
| vivo | We are supportive of the opption3. We want to clarify that in case of single CSI generation model and N multiple CSI reconstruction models working for the truncation of payload (i.e., N truncated payloads), one model ID is only associated with paired single CSI generation model and single CSI reconstruction model (only for one truncated payload). In this case, it requires N model IDs. |
| Ericsson | Add **Option 5**, The pairing ID indicates a reference to a prior training session (e.g. using an API) between NW and UE, associated with a specific dataset. Hence the ID doesn’t refer to a model but instead a process used to obtain the models at each side. |
| Xiaomi | As commended in the last proposal, we should firstly discuss whether the two-sided model is managed by functionality identification or model identification.  In our view, Option 3 is a simple way. |
| LG Electronics | Similar view with proposal 2-3-2. It can be further discussed after some progress in above issue. |
| OPPO | More conclusions on the model/functionality ID issue in 9.2.1 are needed before discussing how to design/associate/use a model ID in CSI part. |
| CMCC | We are generally OK. But for the different definitions of model ID, the feasible option can be different. For example, if the model is logical model, Option 3 is more suitable, while if it is physical model, Option 1 or 2 is also possible. |
| Fujitsu | Support. |
| CAICT | Fine to list all options for CSI configuration and report. It seems that there is similar discussion in 9.2.1 on pairing ID. |
| Futurewei | In general, we think using “pairing ID” would be better. However, this proposal is coupled with the LCM related discussion in 9.2.1 and tied to the previous proposal, thus, we suggest waiting for some progress is made in 9.2.1 and company’s views on the previous proposal. |
| Lenovo | We believe this proposal should be discussed after we have clear view on functionality/mode of LCM. |
| NVIDIA | Support the proposal in principle. |
| ZTE | We should deprioritize this issue. |
| InterDigital | It is better to wait until the definition of model identification and functionality identification becomes clearer under AI 9.2.1 as proposed by few other companies |
| MediaTek | We support the proposal in principle. We think in the option 4, we can add training type 2 as well. |

## Performance monitoring, model update, activation/de-activation/switching

Following table summarize company’s proposals related to model performance monitoring, activation/de-activation/switching.

|  |  |
| --- | --- |
| **Company** | **Key Proposals/Observations/Positions** |
| Huawei | Proposal 16: The input or output data based monitoring should be evaluated at 9.2.2.1 before being further discussed at 9.2.2.2, including: what metrics can be adopted for evaluating the distribution, how to generate the distribution of data, how accurate the data drift reflects the AI/ML model performance.  Observation 9: If monitoring of input data drift is to be further studied, the data drift or out-of-distribution can be reflected by probability distribution function (PDF) or centroids between monitored input data and training data.  Observation 10: Motivation for output data drift is not clear, since the failure of AI/ML model may not be reflected by the output drift.  Observation 11: In CSI compression, if eventual KPI is adopted as monitoring metric, the potential spec impact for methods of removing the impacts of other factors other than model performance   * is up to the Network implementation for the Network side monitoring mode. * can be studied for the UE side monitoring mode.   Proposal 17: To assist the model monitoring by taking into account the aspect of power consumption, it can be considered to introduce the metric report of power consumption from UE to gNB, e.g., whether the power consumption of the undergoing AI/ML model is higher (and how much higher if so) than the legacy non-AI/ML method.  Proposal 18: For Network side monitoring based on intermediate KPI, study the reporting of the target CSI and the associated CSI report by the UE via UCI with higher priority.  Proposal 19: For UE side monitoring based on intermediate KPI, study the indication of the recovery CSI with RRC signaling.   * The association to the corresponding CSI report fed back at the inference stage can be indicated in together.   Proposal 20: For UE side performance monitoring, Network may configure a threshold criterion to facilitate UE to perform model monitoring from the following aspects:   * Usage of the threshold criterion, e.g., UE to perform conditional report of monitoring metrics, or to make the conditional monitoring decisions such as deactivation, switching, etc., based on the threshold. * Types of the threshold criterion, e.g., eventual KPI (e.g., ACK/NACK ratio, throughput, RSRP, etc.) and/or intermediate KPI (e.g., SGCS, NMSE, etc.).   Proposal 21: For the co-existence between AI/ML-based CSI feedback and legacy CSI feedback, further study:   * Configuration/indication of AI/ML-based measurement/report and legacy CSI measurement/report, e.g., configuring separate time durations of different CSI feedback mechanisms, indicating differentiated measurement resources, etc. * Configuration/indication of the precoding type applied to the PDSCH transmission for UE side performance monitoring, e.g., PDSCH precoded by using AI/ML-based CSI feedback or non-AI/ML-based CSI feedback. |
| OPPO | Proposal 6: Regarding the performance monitoring metrics/methods for AI/ML model monitoring, eventual KPIs(e.g., hypothetical BLER) should be utilized for the performance monitoring, other options can be used to equivalent convert the eventual KPI.  Proposal 7: The stability of the performance evaluating and decision-making mechanism should be further studied to avoid the interference of random effects on the evaluation results.   * multiple attempts within an evaluation window both in PHY and high layers would be helpful to obtain a relatively stable evaluation result * multi-user involved mechanism should be addressed |
| vivo | 1. Monitoring inference accuracy is the most direct and reliable performance monitoring method for CSI compression with two-sided models. 2. Legacy codebook with potential enhancement can be used to report CSI measurement for performance monitoring at NW side in CSI compression. 3. Study monitoring inference accuracy at NW side as a baseline for performance monitoring in CSI compression. 4. For NW-side monitoring based on intermediate KPIs, study the necessity and specification impacts of enhancing legacy codebook configurations for CSI measurement reporting. 5. For UE-side monitoring based on the output of the CSI reconstruction model at NW side, study the feasibility and specification impacts of compressing output CSI indication over-the-air. 6. Using system KPIs for performance monitoring in CSI compression might have difficulties in judging whether an observed system performance degradation is caused by an outdated CSI compression model or some other reasons. 7. Monitoring based on data distribution can be viewed as a special case of monitoring based on applicable condition. 8. There could be accuracy and reliability issues for monitoring methods based on applicable condition. 9. Design of applicable condition-based performance monitoring methods and development of scenario-/configuration-/site-specific models should be jointly considered in CSI compression. 10. Study model ID based LCM procedure for CSI compression with two-sided models. 11. Study mechanisms for the two sides to jointly select a model among multiple candidate models, including:  * Triggering conditions * How to conduct multi-model performance monitoring for purpose of model selection * Sharing of model selection results between NW and UE in CSI compression, where model ID based solution can be considered as a starting point.  1. Study the potential specification impact of triggering conditions for Model selection, switching/activation/deactivation, fallback. 2. For ID based model management, study the following options for signaling design for model switching/activation/deactivation among multiple models: RRC-based, MAC CE-based, DCI-based. |
| Spreadtrum Comm | Observation 3: For UE-side performance monitoring, eventual KPIs and input data based monitoring metric can be considered.  Observation 4: For NW-side performance monitoring, eventual KPIs, legacy CSI based monitoring and output data based monitoring metric can be considered. |
| Nokia | Proposal 8: For CSI compression, RAN1 shall study the potential specification impact on performance monitoring by considering   * Methods of performance monitoring (NW-sided, UE-sided, hybrid) * Changes to the reporting framework (e.g., ground-truth reporting to enable performance monitoring at the gNB, KPI reporting when UE considers performance monitoring) * Changes to the measurement framework (e.g., configuring monitoring KPIs and measurement resources) |
| CATT | Proposal 6: In CSI compression using two-sided model use case, for NW side intermediate KPIs based monitoring, further study the signaling and procedures for reporting target CSI, with the following two options considered:   * Option 1: The target CSI is reported together with its associated CSI report; * Option 2: The target CSI is reported separately from its associated CSI report.   Proposal 7: In CSI compression using two-sided model use case, for NW side intermediate KPIs based monitoring, potential specification impact includes the following:   * How to determine the association of target CSI and CSI report by the NW side; * Signaling and procedures for triggering target CSI reporting; * Types of 3GPP signaling takes responsibility on target CSI reporting, e.g., physical signaling, RRC signaling; * Types of target CSI for model monitoring, e.g., precoding matrix, channel matrix etc.; * Formats of target CSI for model monitoring: scaler quantization and/or codebook-based quantization (e.g., e-type II like).   Proposal 8: In CSI compression using two-sided model use case, for UE side intermediate KPIs based monitoring, obtaining the output of the CSI reconstruction model based on the CSI reconstruction model by the UE is only supported for AI/ML model trained with training collaboration Type 1 at UE side.  Proposal 9: In CSI compression using two-sided model use case, for UE side intermediate KPIs based monitoring, further study the signaling and procedures for transmitting output-CSI-UE from NW side to UE side, with the following options considered:   * Option 1: The output-CSI-UE is transmitted to the UE in form of quantization values, e.g., scalar quantization or codebook-based quantization; * Option 2: The output-CSI-UE is transmitted to the UE in form of transmitting precoded CSI-RS that precoded with the output-CSI-UE.   Proposal 10: In CSI compression using two-sided model use case, for UE side intermediate KPIs based monitoring, potential specification impact includes the following:   * How to determine the association of output-CSI-UE and CSI report by the UE; * Signaling and procedures for indicating output-CSI-UE transmission; * Types of 3GPP signaling takes responsibility on transmitting output-CSI-UE, e.g., physical signaling, RRC signaling; * Types of output-CSI-UE for model monitoring, e.g., precoding matrix, channel matrix etc.; * Formats of output-CSI-UE for model monitoring: scaler quantization and/or codebook-based quantization (e.g., e-type II like).   Proposal 11: In CSI compression using two-sided model use case, for UE-side model performance monitoring, study potential specification impacts on the following:   * Content on model performance that UE reports to the network   + Value of monitoring metric;   + Judgement on whether a model is failed, etc. * Signaling/procedure for reporting the performance.   Proposal 12: In CSI compression using two-sided model use case, for UE-side model performance monitoring, if UE side reports the judgement on whether the model is failed to the NW side, study potential specification impact on the criterion on determining whether an AI/ML model is failed or not. |
| NEC | Proposal 3: For UE-side performance monitoring, study how to report the performance metric(s).  Proposal 4: For one AI/ML model of CSI compression, consider monitoring the performances of multiple different ranks.  Proposal 5: Study simultaneous model monitoring for multiple AI/ML models of CSI compression. |
| Intel | Observation 1: Model performance monitoring based on intermediate KPI or eventual KPI calculated based on one AI-ML model is not giving enough information for proper configuration of AI-ML Model  Proposal 1: Testing of different AI-ML models with the measured channel should be considered for model performance monitoring  Proposal 2: For CSI compression using two-sided model use case, further study the necessity, feasibility, and potential specification impact of intermediate KPIs based monitoring for   * NW-side monitoring based on the target CSI from SRS and output CSI obtained from SRS measurements using the two-sided model (assuming that CSI generation part of the model is known at the gNB)   Proposal 3: Co-existence and fallback mechanism between AI/ML-based CSI feedback mode and legacy non-AI/ML-based CSI feedback mode should be based on existing CSI framework   * AI/ML based CSI report can be configured by using CSI-ReportConfig with the corresponding measurements and trigger states (for aperiodic CSI) configuration   Proposal 4: Consider existing NR features as baseline for data collection (e.g., SRS, CSI-RS, CSI reporting)   * The following enhancements can be considered: new codebook design (ground-truth CSI quantization), relaxed timing requirements, reporting of CSI for multiple CSI instances in one CSI report |
| Interdigital | Observation 4: It is possible that the AI/ML encoders do not generalize well across all realistic channel conditions.  Proposal 8: Study means to configure/reconfigure the UE with the monitoring configuration, including the monitoring metric.  Proposal 9: For UE-side monitoring, study triggers and means for reporting the monitoring metrics, including periodic and aperiodic reporting.  Proposal 10: For UE-side monitoring, study appropriate monitoring metrics to avoid unnecessary model updating or switching.  Proposal 11: For UE-side monitoring, study the UE-side monitoring metrics (including report size, metrics quantization, report frequency) to avoid increasing the feedback overhead.  Observation 5: Potential specification impacts of UE-side monitoring based on the output of the (NW-side) CSI reconstruction model include (but may not be limited to): format of the reconstructed CSI, CSI type (full channel matrix or eigenvector), identification of the corresponding CSI report, information on quantization of the reconstructed CSI.  Observation 6: UE-side monitoring based on the output of the NW-side CSI reconstruction model may increase the downlink overhead, because the output CSI reconstructed at the NW needs to be indicated by the NW to the UE.  Observation 7: Potential specification impacts of UE-side monitoring based on the output of the CSI reconstruction model at the UE-side include (but may not be limited to): UE indication of the ID of the UE-side reconstruction model, and means to adjust the intermediate KPI to account for the difference between the UE-side and NW-side CSI reconstruction models.  Observation 8: UE-side monitoring based on the output of the CSI reconstruction model at the UE-side appears to have lower overhead compared to UE-side monitoring based on the output of the NW-side CSI reconstruction model.  Observation 9: NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report has the potential to increase the feedback overheads as the target CSI is reported by the UE or obtained from the UE-side.  Proposal 12: In case of NW-side monitoring, study monitoring approaches with low signaling overhead.  Proposal 14: For CSI compression using two-sided model use case, study AIML model switching or AI/ML model (parameter) update to mitigate AI/ML model performance degradation.  Proposal 15: Study mechanisms for fallback to legacy CSI reporting (e.g. for cases when AIML model performance is poor). |
| Fujitsu | Proposal-3: For the CSI compression using two-sided AI/ML models use case, study the procedures and signaling needed for intermediate KPI-based AI/ML performance monitoring and the follow-up mechanism after the monitoring, including the falling back to codebook-based CSI report from AI/ML-based CSI report.  Proposal-4: For the AI/ML model performance monitoring of the CSI compression using two-sided AI/ML models use case, study the potential specification impacts on monitoring the performance of an AI/ML model in inactivate mode, taking at least the following cases into consideration.   * Initial activation of an AI/ML model. * Re-activation of an AI/ML model. |
| Xiaomi | Proposal 10: In order to improve the reliability of model performance monitoring, legacy CSI based feedback, e.g., eType II-based CSI feedback, should be adopted as a reference. |
| Panasonic | Observation 19: Further study Direction 1 and Direction 3 with proxy model framework.   * Direction 1: Network-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE or obtained from the UE side. * Direction 3: UE-side monitoring based on the output of the CSI reconstruction model at the UE side. |
| Google | Proposal 10: Study the AI/ML model monitoring for CSI compression based on the following options:   * Option 1: NW-based model monitoring, where the performance for the CSI compression is monitored by the gNB and the UE may report some assistant information * Option 2: UE-based model monitoring, where the performance for the CSI compression is monitored by the UE and the UE can report an indication to the NW if it identifies an AI/ML model performance failure   Proposal 11: The metric for AI/ML model monitoring for CSI compression based on the hypothetical BLER measured from precoded CSI-RS with the precoder selected from decompressed CSI in the most recent ML based CSI report.  Proposal 12: Do not support to use SGCS as the metric for ML performance monitoring. |
| LGE | Proposal #8: Consider at least following aspects for fallback operation   * Condition of Fallback mode * NW initiated Fallback mode |
| CAICT | Proposal 3: Original CSI information to be compressed at UE side could be feedback to NW side for model monitoring/training/updating.  Proposal 4: Both periodic and non-periodic original CSI feedback need to be considered.  Proposal 5: NW-side monitoring based on target CSI with realistic channel estimation feedback from UE could be considered as baseline for AI/ML model monitoring.  Proposal 6: UE-side monitoring based on the output of the CSI reconstruction model at the UE-side or indicated by the NW from the network side could be considered as assistant. |
| ETRI | Proposal 3: Consider further studies on potential specification impacts of model selection using the performance monitoring result.  Proposal 4: Consider further studies on potential specification impacts of model changes and fallback operation using the performance monitoring result.  Proposal 5: Study the potential specification impacts on transferring results of CSI prediction.  Proposal 6: Study the potential specification impacts on the level y collaboration for model monitoring.  Observation 4: For AI/ML based CSI prediction, the performance reduction occurs significantly depending on changes of UE speeds and carrier frequency. |
|  |  |
| MediaTek | 1. Study spec impact, signalling requirements, and candidate representative information of AI/ML models for activation, deactivation, switching, and fallback. 2. Discuss methods and apparatus for monitoring AI/ML models other than the one which is already being used by UE and gNB. 3. Given promising features of input/output-based monitoring, accuracy of possible solutions shall be further studied. 4. Prioritize UE-side (Alternative 1) proxy-based model monitoring as the initial monitoring method for tracking intermediate KPI. 5. System-level indicators cannot be regarded as the single point of decisioning for detection of monitoring events. 6. Study multi-stage monitoring approach where a low-overhead low-accuracy method triggers a more accurate intermediate-KPI based solution with higher overhead. 7. Study signalling and ID assignment procedure for AI/ML models generalized over multiple input, output, and latent dimensions. |
| Lenovo | 1. Study the specification impact corresponding to AI model performance monitoring, as well as the corresponding scheme adaptation decision 2. The following four scheme adaptation decisions under AI model performance monitoring are considered as a starting point: (i) No AI model change, (ii) CSI parameters update, (iii) AI model parameter update, (iv) AI model switching, and (v) Fallback to non-AI scheme 3. Fallback to non-AI CSI feedback scheme is considered a part of the scheme adaptation mechanism 4. Network-based performance monitoring and model adaptation are supported by default 5. Further study the specification impact corresponding to the model monitoring schemes: (i) The network configuring the UE to report performance metrics that aid model monitoring, (ii) the network transmitting performance metrics to aid UE-based model monitoring, and (iii) Event-triggered AI model monitoring |
| Qualcomm | Observation 12: Real-time performance monitoring that incurs overhead and/or additional processing complexity is unnecessary.  Observation 13: Model monitoring based on ground-truth provided by UE to the network requires large signaling overhead and may be sensitive to large latency.  Observation 14: Model monitoring using a proxy model that outputs the intermediate KPI directly shows an accurate inference accuracy prediction.  Observation 15: Model monitoring based on metrics derived by comparison between input samples inference and training samples can have strong relationship with the inference accuracy. As a result, input-based monitoring appears promising.  Proposal 13: For model performance monitoring, specification change for reporting the target CSI with high resolution from UE to network requires clear justification as it incurs additional overhead and may not be necessary.  Proposal 13: Study specification impact of methods that directly outputs intermediate KPI at the UE side.  Proposal 14: Study specification impact of input-based model monitoring on the UE-side by comparing input samples at inference time to the training samples. |
| AT&T | Proposal 3: In CSI compression using two-sided model use case, further study potential specification impact needed to enable model performance monitoring using an existing CSI feedback scheme as the reference to compare whether/how much AI/ML outperforms the existing CSI feedback scheme. |
| NTT DOCOMO | Proposal 9: Discuss the feasibility of the model monitoring based on the input/output data distribution in CSI compression, before the specification impact discussion related to it. |

### ***Proposal 2-4-1:***

***In CSI compression using two-sided model use case, for UE-side monitoring, further study potential specification impact on triggering and means for reporting the monitoring metrics, including periodic and aperiodic reporting.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | We suggest we add “UE initiate” before “periodic and aperiodic”. We can define a certain event, if the event happens, the UE can initiate the report, similar to BFR. |
| CATT | Support. |
| Huawei/HiSi | Support in principle. |
| vivo | Support. |
| Xiaomi | Support |
| LG Electronics | Fine with proposal. |
| OPPO | Support. |
| CMCC | Support in principle.  “semi-persistent” can also be added after “including”. |
| NEC | Support |
| Fujitsu | Support in general. |
| CAICT | Support |
| Futurewei | We are fine with the proposal. |
| Lenovo | Support |
| ZTE | Generally fine with the proposal. We suggest rewording it from a high level as  ***In CSI compression using two-sided model use case, for UE-side monitoring, further study potential specification impact on triggering and means for reporting the monitoring metrics~~, including periodic and aperiodic reporting.~~***  since the reporting configuration not only includes periodic and aperiodic reporting, but semi-persistent reporting is also applicable. |
| InterDigital | Support |
| MediaTek | Support |

### ***Proposal 2-4-2:***

***In CSI compression using two-sided model use case, further study potential specification impact needed to enable model performance monitoring and fall back using an existing CSI feedback scheme as the reference to compare whether/how much AI/ML performance compare to the existing CSI feedback scheme.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | It seems this has no spec impact? The NW can configure two CSI report configurations, one for ML, the other for legacy. |
| NTT DOCOMO | Fine with the proposal. |
| CATT | Assuming the concurrent use of AI/ML CSI feedback and non-AI/ML CSI feedback is always guaranteed, we are OK with the proposal.  If not, we should add ‘feasibility’ as part of the study. |
| Huawei/HiSi | Support.  As the AI/ML may suffer model failure in case the environment changes, it is essential for NW and UE to monitor the performance with legacy and trigger/request the fall back on demand. In addition, the comparison metric includes both intermediate KPI (e.g., SGCS) and eventual KPI (e.g., UPT) to provide comprehensive perspectives to identify the model failure.  ***In CSI compression using two-sided model use case, further study potential specification impact needed to enable model performance monitoring and fall back using an existing CSI feedback scheme as the reference to compare whether/how much AI/ML performance compare to the existing CSI feedback scheme***   * ***The metric for monitoring and comparison includes intermediate KPI and eventual KPI.*** |
| vivo | Support. |
| Ericsson | Agree with Google such behaviour is already in the CSI framework. The discussion here is whether and how to enable and disable an AI based CSI compression report, the fallback to legacy is business as usual and already in spec. We object to Autonomous UE fallback within a single configured CSI report |
| Xiaomi | Support. |
| LG Electronics | Similar view with Google. |
| OPPO | Support |
| CMCC | Support. |
| Fujitsu | We suggest revise the wording as “… ***to compare whether/how much the AI/ML performance is beneficial compared to*** …” |
| CAICT | Support |
| Futurewei | We are ok with the proposal in general. However, the first part is more essential (i.e., “***enable model performance monitoring and fall back using an existing CSI feedback scheme”***) as it may have potential spec impact (configuration, triggers, etc.) and using an existing CSI feedback approach as reference to determine the performance of AI/ML based method is just one of the potential uses for this. We suggest removing the second part of the proposal. |
| Lenovo | Agree with Google’s comment. Can we suggest to modify to “***further study potential specification impact, if needed, to enable model performance monitoring and fall back …***” |
| NVIDIA | Support the proposal in principle. |
| ZTE | We think this proposal should be evaluated in 9.2.2.1. |
| InterDigital | Support |
| MediaTek | Support |

### ***Proposal 2-4-3:***

***In CSI compression using two-sided model use case, potential specification impact for input or output data-based monitoring will be further discussed after initial evaluation is performed in 9.2.2.1, including:***

* ***What metrics can be adopted for evaluating the distribution,***
* ***How to generate the distribution of data,***
* ***How accurate the data drift reflects the AI/ML model performance***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | Support |
| NTT DOCOMO | Support |
| CATT | OK. |
| Huawei/HiSi | Support |
| vivo | Support. |
| Ericsson | Support |
| Xiaomi | Support |
| LG Electronics | Fine with proposal. |
| OPPO | Support |
| CMCC | Support. |
| NEC | Support |
| CAICT | Support |
| Futurewei | In general, we are ok. However, we would like to point out that the distribution of the input to the encoder change does not always cause the distribution change of the encoder output (which is a latent representation of the input). Input data distribution change is an issue ONLY if the performance is impacted. Thus, in such case, the potential issue with monitoring the encoder input distribution change is that it may have many false alarms. Monitoring the encoder output distribution change is more relevant to the performance as the encoder output distribution change is more related to CSI reconstruction part’s performance. |
| Lenovo | Support |
| NVIDIA | Support |
| ZTE | Support |
| InterDigital | Support |
| MediaTek | Support |
| Intel | We prefer to start discussion/evaluation in AI 9.2.2.1 directly. |

## Framework, UE capability, and other topics

Following table summarize company’s proposals related to framework.

|  |  |
| --- | --- |
| **Company** | **Key Proposals/Observations/Positions** |
| Huawei | Proposal 22: Study the potential specification impact for UE capability, including the following as a starting point: data collection, dataset delivery, training, model switching, model updating, monitoring, and CSI report timeline. |
| ZTE | Proposal 23: In CSI compression using two-sided model use case, further study potential specification impact of the following UE capability options:   * Framework for defining and reporting UE dynamic capability for model inference. * Whether and how LCM-related procedures are captured into UE capability. |
| Spreadtrum | Proposal 1: Legacy CSI framework can be reused for the sub use case - Spatial-frequency domain CSI compression. Additional enhancement can be considered.  Proposal 2: To facilitate the discussion, views on Pros and Cons of all of Training types are needed to be aligned. What shown in Table 1 can be considered. |
| Nokia | Proposal 1: For the two-sided CSI feedback compression sub-use case, RAN1 shall define applicable conditions for functionalities to enable functionality-based LCM.  Proposal 2: For the two-sided CSI feedback compression sub-use case, RAN1 to study the following applicable conditions for functionalities,  • CSI-RS measurement conditions  • CSI-RS and CSI reports configuration conditions  • CSI calculation conditions (i.e., number of occupied CPUs)  • Output CSI conditions  • Compression ratio conditions (e.g., CR4, CR8, …)  • Quantizer conditions (e.g., SQ1, VQ1, …)  • Pairing ID (e.g., model ID, dataset ID)  • Generic conditions on supporting ML functionalities  Proposal 3: For the two-sided CSI feedback compression sub-use case, UE reports applicable conditions for functionalities by using UE capability reporting.  Proposal 4: For the two-sided CSI feedback compression sub-use case, the NW creates/configures functionalities to the UE with each functionality referring to a configuration message (e.g., RRC) that contains NW-selected applicable conditions (according to the UE capability). |
| CATT | Proposal 13: For AI/ML based CSI feedback, the CSI reporting framework in Rel-17 for codebook based CSI feedback can be reused. |
| NEC | Proposal 1: Support the adjustment of CSI feedback rate/ CSI reporting pattern based on the predicted CSI variation points as a sub-use case of the time-domain CSI prediction. |
| Sony | Proposal 5: RAN1 should study whether the compressed channel information is treated as a new PMI type or new CSI feedback information.  Proposal 6: RAN1 should study specification impact of new PMI type for the CSI compression using two-sided model use case. |
| Xiaomi | Proposal 11: UE side model or UE part model for CSI compression feedback can be identified through AI/ML functionality, AI/ML model, or both functionality and model. How to define AI/ML functionality and/or AI/ML model for CSI compression feedback should be firstly studied. |
| China Telecom | Proposal 1: A new CSI feedback signaling framework design should be standardized based on the legacy non-AI/ML-based CSI feedback mode in Rel-17, e.g., CSI-RS/CSI reporting configurations and CSI processing procedures.  Proposal 2: In CSI compression using two-sided model use case, further study potential specification impact of the UE capability, including the framework for defining and reporting UE capability, model training/updating/monitoring/inference. |
| Google | Proposal 1: The study of AI/ML based CSI compression should be based on the CSI framework in Rel-17.  Proposal 2: The input of CSI compression based on the eigenvectors of the raw channel with a wideband precoder selected as SD basis, e.g. HW1.  Proposal 3: The output of CSI compression should be the compressed eigenvectors for the raw channel with a wideband precoder selected as SD basis, e.g. HW1.  Proposal 4: The CSI report for CSI compression should comprise the beam index(es) for W1 selection and compressed eigenvectors for the raw channel with a wideband precoder selected as SD basis, e.g. HW1.  Proposal 5: If the input of the ML is the frequency domain channel, the UE reports L1-SINR only instead of reporting RI/CQI.  Proposal 6: If the input of the ML is the channel eigenvector or W2, the UE reports a list of CRIs and CQI based on a set of port selection CSI-RS resources.  The gNB applies the decompressed precoders to each CSI-RS resource  Proposal 7: The priority for non-ML based CSI report should be higher than the priority of ML based CSI report.  Proposal 8: The AI/ML based CSI compression should consider the following types of UE:  Type 1 UE (low performance UE): CSI compression is based on general processing unit (GPU)  Type 2 UE (high performance UE): CSI compression is based on neural processing unit (NPU)  Proposal 9: Study the AI/ML model adaptation for CSI compression, where different AI/ML models may be with different compression ratio. |
| Apple | Proposal 3: NW can configure AI based CSI compression with enhanced MIMO related RRC configuration. |
| AT&T | Proposal 4: The study of AI/ML based CSI compression should be based on the legacy CSI feedback signaling framework. Further study the necessity and benefits of any potential specification enhancement for   * CSI-RS configurations * CSI reporting configurations * CSI processing procedures. * Other aspects are not precluded. |

### ***Proposal 2-5:***

***The study of AI/ML based CSI compression should be based on the legacy CSI feedback signaling framework. Further study potential specification enhancement on***

* ***CSI-RS configurations***
* ***CSI reporting configurations***
* ***CSI processing procedures.***
* ***Other aspects are not precluded.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | Support |
| NTT DOCOMO | Fine with the proposal. |
| CATT | Support. |
| Huawei/HiSi | We are OK with the proposal in principle. As in the last meeting we have agreed to discuss the priority/omission rule and CPU (which also impacts the omission); in this meeting, we propose that how to map the AI/ML based CSI feedback to the 2 parts CSI may also have spec impact. Therefore, add one sub-bullet on “CSI report mapping/priority/omission”.   * ***CSI-RS configurations*** * ***CSI reporting configurations*** * ***CSI report mapping/priority/omission*** * ***CSI processing procedures.*** * ***Other aspects are not precluded.*** |
| vivo | Support. |
| Ericsson | OK |
| Xiaomi | Support and we are fine with HW’s updated version. |
| LG Electronics | Fine with proposal. Also fine with HW’s modification. |
| OPPO | Support |
| CMCC | Support. |
| NEC | OK |
| Fujitsu | Support. |
| CAICT | Support |
| Futurewei | We are ok with the proposal and Huawei’s edits. |
| Lenovo | We do not support CSI-RS configuration enhancements, this is out of scope and further studying CSI-RS configuration is not a realistic goal given the time remaining before the conclusion of the study |
| NVIDIA | Support the proposal in principle. |
| ZTE | Agree with Huawei. |
| InterDigital | OK |
| MediaTek | Support |

# Potential specification impact on CSI prediction

Following table summarize company’s proposals related to CSI prediction sub-use case.

|  |  |
| --- | --- |
| **Company** | **View** |
| Huawei | Proposal 1: Start the study of the potential spec impact of CSI prediction after RAN1#112b-e meeting. |
| ZTE | Observation 14: Based on current evaluation assumptions in 9.2.2.1, AI-based CSI prediction shows better performance gain when baseline is the nearest historical CSI. However, AI-based CSI prediction almost has similar performance when the baseline is non-AI/ML based CSI prediction.  Proposal 25: Deprioritize the specification impact discussion on the sub-use case of time domain CSI prediction in Rel-18 SI. |
| OPPO | Proposal 8: For R18 time domain CSI prediction, the two following aspects should be studied to evaluate the performance gain and identify the potential spec impacts.   * Improvement of throughput * Reduction of CSI-RS overhead |
| vivo | Specification impact of AI based CSI prediction should be discussed in R18 AI/ML   1. The model training of AI-based CSI prediction should be discussed with the consideration of NW-side training and UE-side training. 2. For the data collection of historical CSIs, the continuity and sequential order of CSIs in one sample should be guaranteed, which impacts the storage of CSIs and the reporting mode of CSIs to the NW (if needed). 3. Data collection of future CSIs is different for periodic and aperiodic CSI prediction. 4. If data transfer is needed, the delay requirement of data collection differs between model training and monitoring, which may result in different transmission solutions. 5. Data collection of AI-based CSI prediction should be studied. 6. New or combined RS configurations to support the collection of labels if labels are not on the future instances of model input. 7. The assistance information (applicable condition) of collected data for AI based CSI prediction should be configured or reported. 8. The monitoring and a level y/z collaboration-based model adjustment such as model selection/switching, finetuning, deactivation and fall back, are needed to ensure the real time performance of AI-based CSI prediction. 9. Monitoring of AI-based CSI prediction needs to be under the control of NW. 10. Monitoring of AI-based CSI prediction should be studied with the consideration of NW-side calculating and UE-side calculating. 11. The update of applicable condition should be configured/reported after the gNB/UE monitoring. 12. The model adjustment such as model selection/switching, finetuning, deactivation and fallback is essential for CSI prediction to overcome the generalization problem. 13. The decision of model adjustment of AI-based CSI prediction should be controlled by NW. 14. The triggering and signaling to support model adjustment of AI-based CSI prediction should be studied. |
| Spreadtrum | Proposal 16: For CSI predication, the potential specification impact includes model identification, model activation/deactivation/update/switching/monitoring by UE or NW, and assisted signaling such as measurement resources and predicted CSI window. |
| Nokia | Proposal 15: As basic channel prediction scheme report Type II CSI like W1, W2, and Wf for the future time instance tpredict. The AI/ML model of the UE predicts the CSI from N semi-persistent CSI RSs with a repetition rate of, e.g., 5 ms within the observation window of length tobserve.  Proposal 16: Support wideband CSI RS configurations, where all active UEs predicting CSI can observe the radio channel with the widest possible RF bandwidth.  Proposal 17: For high speed UEs consider options to ensure sufficient oversampling for the CSI RS based channel observations as basis for proper channel prediction and generalization UE-sided CSI prediction, RAN1 shall define applicable conditions for functionalities to enable functionality-based LCM.  Proposal 19: For UE-sided CSI prediction, RAN1 to support at least the following applicable conditions for functionalities,   * Support Type II CSI prediction (Supported CSI prediction mode (e.g., TypeII, delay Doppler domain) * Measured CSI RS periodicity (e.g., 5ms, 10ms, 20ms), Prediction time steps (K = 1, 2, 4, [8]), Measured allocation of CSI RS (AE, beam), Measured CSI RS dimension (e.g., 4, 8, 16, 32), Measured CSI RS pattern (e.g., periodic, semi-persistent, aperiodic) * NW-sided performance monitoring conditions (e.g., support measurements of Predicted DL RS set (full Set A, partial Set A), Measurement periodicity (100 ms, 200 ms)) * Conditions on supporting ML functionalities (e.g., Max number of supported functionalities (1, 2, 4, 8,.), Delay on activating a functionality (2 ms, 4 ms), Generalization condition of functionalities (yes, no))   Proposal 20: For UE-sided CSI prediction, RAN1 to study the following additional applicable conditions for functionalities,   * Conditions for UE-sided performance monitoring * Conditions for data collection * Conditions for CSI prediction as predicted time instance versus in the delay Doppler domain * Conditions for assistance info required at the UE like the expected prediction time horizon   Proposal 21: Consider the possibility of overfitting/fine-tuning of UE models for improved CSI prediction. |
| NEC | Proposal 6: Study discontinuous periodic or semi-persistent CSI report.  Proposal 7: Support the location/CQI report timing set mapping table based on AI/ML.  Proposal 8: Support the location/CQI periodicity mapping table based on AI/ML. |
| Intel | Proposal 9:   * CSI prediction with AI/ML model at the UE side shall be discussed in application to Rel-18 CSI enhancements for high/medium mobility   Proposal 10:   * Study model performance monitoring based on intermediate metrics (e.g., GCS) calculated from the measured CSI-RS and predicted channel at the UE side |
| Interdigital | Proposal 19: Specification impact for time domain CSI prediction using UE sided model is not studied in Rel-18. |
| xiaomi | Proposal 12: The specification impact on time domain CSI prediction using UE sided model selected as a representative sub-use case for CSI enhancement could be studied in Rel-18.  Proposal 13: The number of CSI-RS resources and the interval of adjacent CSI-RS resources discussed in Rel-18 MIMO CSI enhancement for medium/high velocities should be as a starting point to study its potential specification impact. |
| China Telecom | Observation 1: Based on current evaluation assumptions in 9.2.2.1, AI-based CSI prediction shows better performance gain when baseline is the nearest historical CSI. However it is not clear how much performance gain can be obtained when baseline is non-AI/ML algorithm based CSI prediction (AR, linear filtering, etc ), and need FFS.  Proposal 3: In CSI prediction using one-sided model use case, further study potential specification impact of the following UE-side training case and NW-side training case:   * Case 1: Both training and inference at UE-side without model transfer * Case 2: Training at NW-side and inference at UE-side with model transfer, e.g., the model structure, model parameters, etc.   Proposal 4: Further study potential specification impact of the procedure of NW-side training and UE-side training based CSI prediction, including data transfer, model transfer, monitoring and adjustments.  Proposal 5: For the UE based CSI prediction, potential specification impact including UE capability signalling, NW and UE’s alignment on prediction related time domain configuration information. |
| Google | Proposal 13: Study the following output of CSI prediction:   * Predicted RI/PMI based on Type1 codebook * Predicted CSI dwelling time |
| LGE | Proposal #9: Study potential specification impacts on UE-sided CSI prediction including at least followings   * AI/ML model monitoring procedure/metric, * enhancement of CSI reporting. |
| CMCC | Proposal 10: For CSI prediction, regarding the spec impact during inference phase, we could take the agreements achieved in Rel-18 9.1.2 sub-agenda as a starting point.  Proposal 11: For CSI prediction, Some CSI related parameters agreed in 9.1.2 sub-agenda might need revision to adapt AI/ML-enabled CSI prediction.  Proposal 12: For CSI prediction, regarding the LCM related potential specification impact, we could take the UE-sided model related agreements achieved in Rel-18 9.2.3.2 sub-agenda as a starting point. |
| MediaTek | 1. For AI/ML-based CSI prediction, discuss the potential specification impact of CSI feedback mechanism. Codebook-based feedback can be used as a baseline (legacy codebook or Rel-18 codebook), but AI/ML-based CSI compression is not precluded. 2. For AI/ML-based CSI prediction, discuss the potential specification impact of the signalling between UE and gNB. |
| Apple | Proposal 10: For CSI prediction use case, potential specification impact including UE capability signaling, UE request and NW activation/de-activation signaling. |
| Lenovo | 1. For AI-based CSI prediction, the baseline for comparison is based on (i) multiple realizations of Rel-16 eType-II codebook-based CSI feedback, and (ii) MIMO Rel-18 codebook design outline 2. CSI feedback for AI-based CSI prediction should follow the same format as legacy CSI feedback in terms of the spatial domain and frequency domain transformations 3. For observation window and prediction window in AI-based CSI prediction, reuse the definitions agreed in Rel-18 MIMO CSI enhancements for high speed 4. Three intermediate KPI values are considered for CSI prediction sub-use case: (i) at the first slot of the prediction window, (ii) at the median slot of the prediction window, and (iii) at the last slot of the prediction window |
| AT&T | Proposal 5: Resume the specification impact discussion for the CSI prediction using UE sided model.  Proposal 6:   * In CSI prediction using UE sided model use case, study the necessity, feasibility, and potential specification impact of UE side data collection enhancement including at least   + Additional CSI configuration information   + Assistance information for UE data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc.     - The provision of assistance information needs to consider feasibility of disclosing proprietary information to the other side.   + Signaling for triggering the data collection * In CSI prediction using UE sided model use case, study the necessity, feasibility, and potential specification impact for NW side data collection including at least:   + Additional CSI configuration information   + Contents of the ground-truth CSI including:     - Data sample type, e.g., precoding matrix, channel matrix etc.     - Data sample format: scaler quantization and/or codebook-based quantization (e.g., e-type II like).     - Assistance information (e.g., time stamps, and/or cell ID, Assistance information for Network data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc., and data quality indicator)   + Latency requirement for data collection   + Signaling for triggering the data collection   Proposal 7: For CSI prediction using UE sided model study the following configurations and their granularity that will be signaled through the functionality, and the corresponding specification impact in functionality-based LCM   * UE speed * Frequency PRB’s * Prediction window * Observation window * Scenario (Uma etc.) * Performance requirement/monitoring * Other additional configurations   Proposal 8: For CSI prediction using UE sided model, study the requirement for model-ID based LCM to support additional functionalities that will not be supported through functionality-based LCM. |
| Samsung | Proposal 1-1: Study the specification impacts of UE-side time-domain CSI prediction under network-UE collaboration level y.  Proposal 1-2: For the AI/ML based CSI prediction sub-use case, study the necessity and specification impact of   * CSI measurement and reporting framework enhancement. * LCM assistance from gNB including, model monitoring, dataset collection, model activation, model deactivation, model switching, etc. |

### ***Proposal 3-1:***

***In CSI prediction using UE-side model use case, start the study of the potential spec impact of CSI prediction after RAN1#112b-e meeting.***

Please provide your view below:

|  |  |
| --- | --- |
| **Company** | **View** |
| Google | OK. Now we already supported CSI prediction based on an enhanced eType2 CSI codebook in MIMO agenda, we think the CSI prediction could focus more on Type1 codebook and CSI dwelling time prediction. |
| CATT | If most of the potential spec impacts for CSI compression are already identified, we are fine to start discussion on CSI prediction case. |
| Huawei/HiSi | We can support the proposal. |
| vivo | We are supportive of the proposal.  The EVMs have already shown benefit of AI-based CSI prediction from both the intermediate and eventual KPIs. Therefore, it is appropriate to start the study of the potential spec impact of CSI prediction after RAN1#112b-e meeting.  In our viewpoint, the definition, configuration of observation window, prediction window and the reporting of predicted CSI can reuse the agreement in Rel-18 MIMO CSI enhancements for high speed.  For AI-based CSI prediction, the specification aspect on the following 3 aspects are prominent:   * data collection procedure, mainly including RS configuration, signalling for alignment of dada collection configuration between NW and UE. * monitoring procedure and metric for AI-based CSI prediction. * model adjustment (model selection/switching and finetuning) procedure.   Other LCM related potential specification impact can follow the high- level principle of general framework and other one-sided model sub-cases.  Thanks to the Rel-18 MIMO WI and the study in other one-sided sub-cases, the workload of AI-based CSI prediction has already been reduced significantly, and the remaining part can be studied from next meeting. |
| Ericsson | Support |
| Xiaomi | We are fine and prefer to firstly discussion on the specification impact for CSI compression. |
| LG Electronics | Fine with proposal. |
| OPPO | Support |
| CMCC | Support. |
| Fujitsu | Support. |
| CAICT | Support |
| Futurewei | We are ok with the proposal. |
| Lenovo | We are fine with the proposal. However, due to the very limited time remaining to study spec impact of CSI prediction (only two meetings after RAN1#112bis-e), **we suggest the following:**  1. The aspects to be studied as part of spec impact discussion are determined in this meeting  2. We have different views from Google regarding the scope of CSI prediction. We strongly prefer reusing the Rel-18 eType-II framework for CSI prediction, which can help significantly reduce the workload needed towards the study. There isn’t enough time to discuss a new framework |
| NVIDIA | Support. |
| InterDigital | No. We have a strong concern on increasing any scope considering current AI/ML overall progress is way behind the schedule. We only finished 30% of study item and just 3 meetings left including this meeting to finalize the study. |
| MediaTek | Support |

# Appendix: Companies input on training collaboration type comparison table.

***Huawei:***

Table 2 Brief comparison of the training types for two-sided model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Type 1 | | Type 2 | Type 3 | |
|  | NW-sided | UE-sided |  | NW first | UE first |
| Whether model can be kept proprietary | No | No | Yes | Yes | Yes |
| Whether require privacy-sensitive dataset sharing | No | No | No | No | No |
| Flexibility to support cell/site/scenario/configuration specific model | Yes | No | No | Yes | No |
| Whether gNB/device specific optimization is allowed | Restricted | Restricted | Yes | Yes | Yes |
| Overhead | Model | Model | N/A | Dataset | Dataset |
| Model update flexibility after deployment | Flexible | Not flexible | Not flexible | Semi-flexible | Not flexible |
| Engineering isolation (feasibility of allowing UE side and NW side to develop/update models separately) | Non-isolable | Non-isolable | Strongly non-isolable | Isolable | Isolable |
| Model performance | Suboptimal | Suboptimal | Suboptimal | Suboptimal | Suboptimal |
| Whether gNB can maintain/store a single/unified model | No | No | Yes | Yes | Restricted |
| Whether UE device can maintain/store a single/unified model | Yes | Yes | Yes | Yes | Yes |
| Extendibility | Support | Restricted | Support | Support | Support |
| Whether training data distribution can match the inference device | Restricted | Yes | Restricted | Restricted | Yes |
| Software/hardware compatibility (Whether device capability can be considered for model development) | Compatibility issue exists | Compatibility issue exists | Free of compatibility issue | Free of compatibility issue | Free of compatibility issue |

***ZTE:***

Table 1. Brief comparison of the training types for two-sided model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training types  Characteristics | Type 1 | | Type 3 | |
| NW side | UE side | NW-first | UE-first |
| Whether model can be kept proprietary | No | No | Yes | Yes |
| Requirements on privacy-sensitive dataset sharing | No | No | No | No |
| Flexibility to support cell/site/scenario/configuration specific model | Yes | No | Yes | No |
| gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware | Restricted (only when the model structure of UE-part model is known by network) | Restricted (only when the model structure of NW-part model is known by network) | Yes | Yes |
| Model update flexibility after deployment | Flexible | Inflexible | Semi-flexible | Inflexible |
| Feasibility of allowing UE side and NW side to develop/update models separately | Infeasible | Infeasible | Feasible | Feasible |
| Model performance based on evaluation in 9.2.2.1 | Optimal | Optimal | Sub-optimal | Sub-optimal |
| Whether gNB can maintain/store a single/unified model | Yes | No | Yes | No |
| Whether UE device can maintain/store a single/unified model | No | Yes | No | Yes |
| Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use | Non-extendable (possible only when UE knows NW-part model in use) | Non-extendable  (possible only when NW knows UE-part model in use) | Extendable | Extendable |
| Whether training data distribution can be matched to the device that will use the model for inference | No for UE device | Yes for UE device | Possible for UE device only when ground-truth CSI is shared from target UE device. | Possible for UE device only when ground-truth CSI is shared from target gNB device |
| Whether device capability can be considered for model development | Yes (when the supported UE-part model structure is known by network side) | UE-part model is up to UE implementation | UE-part model is up to UE implementation | UE-part model is up to UE implementation |

***OPPO***

|  |  |  |
| --- | --- | --- |
|  | Training collaboration type 1 | Training collaboration type 3 |
| Whether model can be kept proprietary | NO  NW/UE needs to deliver a model or a sub-model to the other side | YES  Model or sub-model delivery is not needed |
| Requirements on privacy-sensitive dataset sharing | NO  Training data transmission is not needed in training collaboration type 1 | YES, if CSI training data belongs to privacy-sensitive dataset  Training data transmission is the fundamental requirement of training collaboration type 3.  NW may need to transmit cell level CSI training data to a given UE.  UE may need to transmit UE level CSI training data to NW  FFS whether the CSI training data belongs to privacy-sensitive dataset |
| Flexibility to support cell/site/scenario/configuration specific model | Support  For NW side training: YES  Cell/site/scenario/configuration specific model could be supported and delivered from NW to UE  For UE side training: Also feasible, if there is some information identifying such kind of cell/site/scenario/ configuration for the data collection | Support  For NW first training: YES  Cell/site/scenario/configuration specific training data could be supported and delivered from NW to UE  For UE first training: Also feasible, if there is some information identifying such kind of cell/site/ scenario/configuration for the data collection |
| gNB/device specific optimization | Support, under condition.  For example, (1) NW/UE prepares models that match different devices. For example, when some devices are more compatible with transformer type models, while others require deployment of other model structures (CNN or fully connected networks), model providers must prepare multiple models to meet device specific optimization requirements. Alternatively, (2) NW/UE does not directly use the obtained decoder/encoder, but instead regenerates training data for matching devices based on the obtained model. | Support.  For example, UE can achieve device specific optimization based on the obtained training data and its capabilities. |
| Model update flexibility after deployment | Support  NW can send new models to UE to update the UE side model, or vice versa | Support  NW can send new training data to UE to update the UE side model, or vice versa |
| Feasibility of allowing UE side and NW side to develop/update models separately | Support  For example, after deploying Model 1 on the UE side, a new UE model can be obtained by using Model 1 as the teacher model and using knowledge distillation method | Support  For example, UE can retrain new models based on the obtained training data |
| Model performance based on evaluation in 9.2.2.1 | No extra performance loss caused by mismatched two sided models and separate training procedures. | May cause extra performance loss.  The performance of training depends on the quality and quantity of the transmitted training date. |
| Whether gNB can maintain/store a single/unified model | For NW side training: NW needs to store multiple models, especially when it needs to transmit different models to UE according to different scenarios | For NW first training: NW needs to store multiple models as well as datasets corresponding to multiple models, especially when the network needs to transmit different datasets to help UE complete type 3 training in different scenarios |
| Whether UE device can maintain/store a single/unified model | For UE side training: UE needs to store multiple models, especially when it needs to transmit different models to NW according to different scenarios | For UE first training: UE needs to store multiple models and datasets corresponding to multiple models, especially when the network needs to transmit different datasets to help NW complete type 3 training in different scenarios |
| Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use | Support  As discussed above, after deploying Model 1 on the UE side, the update of the UE model can be achieved by using Model 1 as a teacher model and using knowledge distillation to obtain a new UE model | Support  As discussed above, UE can retrain a new model by the obtained training data |
| Whether training data distribution can be matched to the device that will use the model for inference | Support, under condition.  For NW side training: If UE directly uses the model passed to UE by NW, there is no guarantee that training data distribution can be matched to the device that will use the model for inference.  Unless (1) NW prepares models that match to different devices, or (2) UE does not directly use the obtained encoder locally, but instead regenerates training data for matching devices based on the obtained encoder.  For UE side training: vice versa | Support, under condition.  For NW side training: If UE directly uses the data passed to UE by NW, there is no guarantee that the training data distribution can be matched to the UE that will use the model for inference.  Unless (1) NW prepares different training data that match to different devices, or (2) UE does not directly use the obtained training data locally, but instead adjusts/ regenerates training data for matching devices based on the obtained training data sets.  For UE side training: vice versa |
| Whether device capability can be considered for model development | Support, under condition.  NW/UE needs to prepare different models that match the capabilities of different devices. | Support  NW/UE can construct and train local models based on received training data and their own abilities to complete Type 3 training. However, due to mismatched CSI encoders and decoders, it may result in extra performance loss. |

|  |  |  |
| --- | --- | --- |
|  | Training collaboration type 1 | Training collaboration type 3 |
| Whether model can be kept proprietary | NO | YES |
| Requirements on privacy-sensitive dataset sharing | NO | YES |
| Flexibility to support cell/site/scenario/configuration specific model | Support | Support |
| gNB/device specific optimization | Support, under condition. | Support. |
| Model update flexibility after deployment | Support | Support |
| Feasibility of allowing UE side and NW side to develop/update models separately | Support | Support |
| Model performance based on evaluation in 9.2.2.1 | No extra performance loss | May cause extra performance loss |
| Whether gNB can maintain/store a single/unified model | For NW side training: YES | For NW first training: YES |
| Whether UE device can maintain/store a single/unified model | For UE side training: YES | For UE first training: YES |
| Extendability | Support | Support |
| Whether training data distribution can be matched to the device that will use the model for inference | Support, under certain condition. | Support, under certain condition. |
| Whether device capability can be considered for model development | Support, under certain condition. | Support. |

***vivo***

To summarize our comments regarding this issue, we provide the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| General aspects | Detailed issues | Type1 | Type2 | Type3 |
| Performance | Model performance based on evaluation in 9.2.2.1 | optimal | Near-Optimal | Suffer from losses in some cases. Near-optimal in some other cases. |
| Proprietary issues (model and data) | Whether model can be kept proprietary | No | Yes | Information on model structure may be required to disclose. |
| Requirements on privacy-sensitive dataset sharing | No concerns | No concerns | Concerns on disclosing data from one user to another one. |
| Flexibility issues (model update and engineering separation) | Flexibility to support cell/site/scenario/configuration specific model | Good. New model can be flexibly transferred to UE when UE enters a new cell/site/scenario/ etc. | Not good, since setting up a new training session is required to obtain a new model for the current cell/site/scenarios etc. | Not good, since setting up a new training session is required to obtain a new model for the current cell/site/scenarios etc. |
| Model update flexibility after deployment | Good with re-transferring updated model | Not good, since setting up a new model training session with exchanging FP/BP information is required. | Not good, since setting up a new separate training session is required. |
| Extendibility to multi-vendor configuration/ Engineering isolation | Extendibility: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use | Support by solely training an encoder compatible with existing decoders (and potential other encoders) at a single entity\* | Support by solely training an encoder compatible with existing decoders (and potential other encoders) via FP/BP exchange\* | Support by sending input/output data to the newly arrived UE’s encoder |
| Feasibility of allowing UE side and NW side to develop/update models separately | Not Support | Support | Support |
| Whether gNB can maintain/store a single/unified model. | Support by training common decoder for multiple encoders at a single entity | Support by training common decoder for multiple encoders via FP/BP exchange | Support by training common decoder via collecting data from multiple UEs |
| Whether UE device can maintain/store a single/unified model | Support by training common encoder for multiple decoders at a single entity | Support by training common encoder for multiple decoders via FP/BP exchange | Support by training common encoder via collecting data from multiple gNBs |
| Support of device specific models | gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware | Support when devices report their supported model designs | Support | Support device-specific model design. Not support device-specific data distribution in NW-first training. |
| Whether training data distribution can be matched to the device that will use the model for inference | Support when devices report its data to the training entity | Support | Support when device reports its data to the “first training entity” |
| Whether device capability can be considered for model development | Support with device capability reporting | Support | Support |

***Spreadtrum Comm***

Table 1 Analysis on Pros and Cons of Training type 1, 2 and 3

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Training Type 1** | **Training Type 2** | **Training Type 3** |
| Whether model can be kept proprietary |  |  |  |
| Requirements on privacy-sensitive dataset sharing |  |  |  |
| Flexibility to support cell/site/scenario/configuration specific model |  |  |  |
| gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware |  |  |  |
| Model update flexibility after deployment |  |  |  |
| feasibility of allowing UE side and NW side to develop/update models separately |  |  |  |
| Model performance based on evaluation in 9.2.2.1 | good | well | well |
| Whether gNB can maintain/store a single/unified model | , if gNB generates the model; otherwise |  |  |
| Whether UE device can maintain/store a single/unified model | , if UE generates the model; otherwise |  |  |
| Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use |  |  |  |
| Whether training data distribution can be matched to the device that will use the model for inference |  |  |  |
| Whether device capability can be considered for model development |  |  |  |

***Ericsson:***

**Summary**

To summarize the discussion around Type 1 training, we provide the following table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Aspect | Type 1a-i | Type 1a-ii | Type 1a-iii | Type 1a-iv | Type 1b-i | Type 1b-ii | Type 1b-iii | Type 1b-iv |
| Proprietary models | No | No | No | No | No | No | No | No |
| Requires dataset sharing | Yes | Yes | Yes | Yes | No | No | No | No |
| Support site specific models | Yes | Yes | Yes | Yes | No | No | No | No |
| gNB specific hardware optimization | Yes | Yes | Yes | Yes | No | Yes | Low | No |
| UE specific hardware optimization | No | Yes | Low | No | Yes | Yes | Yes | Yes |
| Allows single unified model on gNB side | Yes | Yes | Yes | Yes | No | No | No | No |
| Allows single unified model on UE side | No | No | No | No | Yes | Yes | Yes | Yes |
| Model update flexibility after deployment | Yes | Delayed and Partial | Partial | Yes | Yes | Delayed and Partial | Partial | Yes |
| Engineering isolation | Low | Low | Low | Low | Low | Low | Low | Low |
| Supports UE-proprietary input | No | No | No | No | Yes | Yes | Yes | Yes |
| Supports NW-proprietary input | Yes | Yes | Yes | Yes | No | No | No | No |
| Supports NW-proprietary output | Yes | Yes | Yes | Yes | No | No | No | No |
| Matching data distribution | No | No | No | No | Yes | Yes | Yes | Yes |
| Capability consideration | Maybe | Maybe | Maybe | Maybe | Yes | Yes | Yes | Yes |
| **Overall feasibility** | No | No | Unclear | Maybe long-term | No | No | No | No |

**Summary**

To summarize the discussion around Type 2 training, we provide the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Aspect | Type 2a-i | Type 2a-ii | Type 2b-i | Type 2b-ii |
| Proprietary models | Yes | Yes | Yes | Yes |
| Requires dataset sharing | Yes | Partially | Yes | Partially |
| Support site specific models | No | No | No | No |
| gNB specific hardware optimization | Yes | Yes | Yes | Yes |
| UE specific hardware optimization | Yes | Yes | Yes | Yes |
| Allows single unified model on gNB side | Yes | Yes | Yes | Yes |
| Allows single unified model on UE side | No | No | Yes | Yes |
| Model update flexibility after deployment | No | No | No | No |
| Engineering isolation | No | No | No | No |
| Supports UE-proprietary input | No | Yes | No | Yes |
| Supports NW-proprietary input | Maybe | No | No | No |
| Supports NW-proprietary output | Yes | Yes | No | No |
| Extendibility | No | No | No | No |
| Matching data distribution | No | Yes | No | Yes |
| **Overall feasibility** | No | No | No | No |

**Summary**

To summarize the discussion around Type 3 training, we provide the following table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Aspect | Type 3a-i-1 | Type 3a-i-2 | Type 3a-i-3 | Type 3a-ii | Type 3b-i-1 | Type 3b-i-2 | Type 3b-ii-1 | Type 3b-ii-2 |
| Proprietary models | Yes | No | No | Yes | Partially | Yes | No | No |
| Requires dataset sharing | Partially | Partially | Partially | Partially | Partially | Partially | Partially | Partially |
| Support site specific models | Yes | Yes | Yes | No | No | No | Partially | Partially |
| gNB specific hardware optimization | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| UE specific hardware optimization | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Allows single unified model on gNB side | Yes | Yes | Yes | Yes | No | No | No | No |
| Allows single unified model on UE side | No | No | No | No | Yes | Yes | Yes | Yes |
| Model update flexibility after deployment | Partial | Partial | Partial | Partial | Partial | Partial | Partial | Partial |
| Engineering isolation | Moderate | Moderate | Moderate | Moderate | Moderate | Moderate | Moderate | Moderate |
| Supports UE-proprietary input | No | No | No | Yes | Yes | Yes | No | No |
| Supports NW-proprietary input | Yes | Yes | Yes | Yes | No | No | Yes | Yes |
| Supports NW-proprietary output | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Matching data distribution | No | No | No | Yes | Yes/No | Yes/No | Yes/No | Yes/No |
| Capability consideration | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Overall feasibility** | Yes | No | No | Yes | No | No | No | No |

**Summary**

To summarize the discussion around Type 4 training, we provide the following table.

|  |  |
| --- | --- |
| Aspect | [Type 4] |
| Proprietary models | Yes |
| Requires dataset sharing | Partially |
| Support site specific models | Yes |
| gNB specific hardware optimization | Yes |
| UE specific hardware optimization | Yes |
| Allows single unified model on gNB side | Yes |
| Allows single unified model on UE side | No |
| Model update flexibility after deployment |  |
| Engineering isolation | Moderate |
| Supports UE-proprietary input | Yes |
| Supports NW-proprietary input | No |
| Supports NW-proprietary output | Yes |
| Matching data distribution | Yes |
| Capability consideration | Yes |
| **Overall feasibility** | Yes |

***Xiaomi:***

Table 1: The pros and cons of joint training of two-sided model at NW sided and NW-first separate training

|  |  |  |
| --- | --- | --- |
| Items | Type 1  (Joint training of the two-sided model at NW side) | Type 3  (NW-first separate training) |
| *Whether model can be kept proprietary* | NO | YES |
| *Requirements on privacy-sensitive dataset sharing* | NO | NO |
| *Flexibility to support cell/site/scenario/configuration specific model* | YES | YES (Depends on UE capability of training AI/ML model) |
| *gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware* | NO（Only NW -sided hardware optimization ） | YES |
| *Model update flexibility after deployment* | YES | YES |
| *Feasibility of allowing UE side and NW side to develop/update models separately* | NO | YES |
| *Model performance based on evaluation in 9.2.2.1* | Optimization | Less than Type I |
| *Whether gNB can maintain/store a single/unified model* | YES | YES |
| *Whether UE device can maintain/store a single/unified model* | NO | NO |
| *Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use* | NO | YES |
| *Whether training data distribution can be matched to the device that will use the model for inference* | Depends on the collected training data | Depends on the collected training data |
| *Whether device capability can be considered for model development* | YES | YES |

***Apple***

**Table I: Comparison of different training collaboration**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Type 1 | | Type 2 | Type 3 | |
|  | UE side | NW side |  | UE first | NW first |
| Model proprietary | No | No | Yes | Yes | Yes |
| Require privacy data sharing | No | UE share dataset to NW | UE share target CSI to NW | UE share target CSI and encoder output to NW | UE share dataset to NW. NW share encoder training dataset |
| Flexibility to support config specific model | Yes, with NW assisted info | Yes | No. | Yes, with NW assisted info | Yes |
| Model upgrade flexibility | Flexible | Most flexible | Not flexible | Flexible | Flexible |
| Develop separately | Yes | Yes | No | Yes | Yes |
| Single model at gNB | No | Yes | Difficult | Yes | Yes |
| Single model at UE | Yes | Yes via frequent download | Difficult | Yes | Performance might degrade |
| Extendibility | Limited | Limited | Limited | Limited | Limited |
| Device specific training data | Yes | Yes, with UE assisted info | Yes | Yes | Yes, with UE assisted info |
| Device capability considered in training | Yes | Yes | Yes | Yes | Yes |

# Appendix: Previous meeting agreements

## RAN1 #109e

Agreement

Spatial-frequency domain CSI compression using two-sided AI model is selected as one representative sub use case.

* + Note: Study of other sub use cases is not precluded.
  + Note: All pre-processing/post-processing, quantization/de-quantization are within the scope of the sub use case.

Conclusion

* Further discuss temporal-spatial-frequency domain CSI compression using two-sided model as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion.
* Further discuss improving the CSI accuracy based on traditional codebook design using one-sided model as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion.
* Further discuss CSI prediction using one-sided model as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion
* Further discuss CSI-RS configuration and overhead reduction as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion
* Further discuss resource allocation and scheduling as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion
* Further discuss joint CSI prediction and compression as a possible sub-use case for CSI feedback enhancement after evaluation methodology discussion.

## RAN1 110

Agreement

In CSI compression using two-sided model use case, the following AI/ML model training collaborations will be further studied:

* Type 1: Joint training of the two-sided model at a single side/entity, e.g., UE-sided or Network-sided.
* Type 2: Joint training of the two-sided model at network side and UE side, respectively.
* Type 3: Separate training at network side and UE side, where the UE-side CSI generation part and the network-side CSI reconstruction part are trained by UE side and network side, respectively.
* Note: Joint training means the generation model and reconstruction model should be trained in the same loop for forward propagation and backward propagation. Joint training could be done both at single node or across multiple nodes (e.g., through gradient exchange between nodes).
* Note: Separate training includes sequential training starting with UE side training, or sequential training starting with NW side training [, or parallel training] at UE and NW
* Other collaboration types are not excluded.

**Conclusion**

CSI-RS configuration and overhead reduction is NOT selected as one representative sub-use case for CSI feedback enhancement use case.

**Conclusion**

Resource allocation and scheduling is NOT selected as one representative sub-use case for CSI feedback enhancement use case.

Agreement

In CSI compression using two-sided model use case, further study potential specification impact on CSI report, including at least

* CSI generation model output and/or CSI reconstruction model input, including configuration(size/format) and/or potential post/pre-processing of CSI generation model output/CSI reconstruction model input.
* CQI determination
* RI determination

Agreement

In CSI compression using two-sided model use case, further study potential specification impact on output CSI, including at least

* Model output type/dimension/configuration and potential post processing

Agreement

In CSI compression using two-sided model use case, further discuss at least the following aspects, including their necessity/feasibility/potential specification impact, for data collection for AI/ML model training/inference/update/monitoring:

* Assistance signaling for UE’s data collection
* Assistance signaling for gNB’s data collection
* Delivery of the datasets.

## RAN1 #110bis-e

Conclusion

Joint CSI prediction and CSI compression is NOT selected as one representative sub-use case for CSI feedback enhancement use case.

Conclusion

CSI accuracy enhancement based on traditional codebook design is NOT selected as one representative sub-use case for CSI feedback enhancement use case.

Conclusion

Temporal-spatial-frequency domain CSI compression using two-sided model is NOT selected as one representative sub-use case for CSI enhancement use case.

• Up to each company to report whether past CSI is used as model input for spatial-frequency domain CSI compression

***Agreement***

In CSI compression using two-sided model use case, study potential specification impact for performance monitoring including:

* NW-side performance monitoring: NW monitors the performance and make decisions of model activation/ deactivation/updating/switching
* UE-side performance monitoring: UE monitors the performance and reports to Network, NW makes decisions of model activation/ deactivation/updating/switching

***Agreement***

In CSI compression using two-sided model use case, further study potential specification impact related to assistance signaling and procedure for model performance monitoring***.***

***Agreement***

In CSI compression using two-sided model use case, further study potential specification impact related to potential co-existence and fallback mechanisms between AI/ML-based CSI feedback mode and legacy non-AI/ML-based CSI feedback mode.

***Agreement***

In CSI compression using two-sided model use case, further study at least the following options for performance monitoring metrics/methods:

* Intermediate KPIs as monitoring metrics (e.g., SGCS)
* Eventual KPIs (e.g., Throughput, hypothetical BLER, BLER, NACK/ACK).
* Legacy CSI based monitoring: schemes using additional legacy CSI reporting
* Other monitoring solutions, at least including the following option:
  + Input or Output data based monitoring: such as data drift between training dataset and observed dataset and out-of-distribution detection

***Agreement***

In CSI compression using two-sided model use case, further study at least use cases of the following potential specification impact on quantization method alignment between CSI generation part at UE and CSI reconstruction part at gNB:

* Alignment of the quantization/dequantization method and the feedback message size between Network and UE

## RAN1 #111

**Agreement**

Time domain CSI prediction using UE sided model is selected as a representative sub-use case for CSI enhancement.

Note: Continue evaluation discussion in 9.2.2.1.

Note: RAN1 Defer potential specification impact discussion at 9.2.2.2 until the RAN1#112b-e, and RAN1 will revisit at RAN1#112b-e whether to defer futher till the end of R18 AI/ML SI.

Note: LCM related potential specification impact follow the high level principle of other one-sided model sub-cases.

Conclusion

In CSI compression using two-sided model use case, training collaboration type 2 over the air interface for model training (not including model update) is deprioritized in R18 SI.

Note:

* To align terminology, output CSI assumed at UE in previous agreement will be referred as output-CSI-UE.
* To align terminology, input-CSI-NW is the input CSI assumed at NW

## RAN1 #112

*Agreement*

In CSI compression using two-sided model use case, further study potential specification impact of the following output-CSI-UE and input-CSI-NW at least for Option 1:

* Option 1: Precoding matrix
  + 1a: The precoding matrix in spatial-frequency domain
  + 1b: The precoding matrix represented using angular-delay domain projection
* Option 2: Explicit channel matrix (i.e., full Tx \* Rx MIMO channel)
  + 2a: raw channel is in spatial-frequency domain
  + 2b: raw channel is in angular-delay domain
* Note: Whether Option 2 is also studied depends on the performance evaluations in 9.2.2.1.
* Note: RI and CQI will be discussed separately

Agreement

In CSI compression using two-sided model use case, further study the following options for CQI determination in CSI report, if CQI in CSI report is configured.

* Option 1: CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation, including
  + Option 1a: CQI is calculated based on target CSI with realistic channel measurement
  + Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment
  + Option 1c: CQI is calculated based on legacy codebook
* Option 2: CQI is calculated based on the output of CSI reconstruction part from the realistic channel estimation, including
  + Option 2a: CQI is calculated based on CSI reconstruction output, if CSI reconstruction model is available at the UE and UE can perform reconstruction model inference with potential adjustment
    - Note: CSI reconstruction part at the UE can be different comparing to the actual CSI reconstruction part used at the NW.
  + Option 2b: CQI is calculated using two stage approach, UE derive CQI using precoded CSI-RS transmitted with a reconstructed precoder.
* Other options are not precluded
* Note1: feasibility of different options should be evaluated
* Note2: Gap analyses between the UE side CQI calculation results and the NW side results, as well as the impact on the scheduling performance should be evaluated
* Note3: Complexity of CQI calculation needs to be evaluated, including the computing complexity and potential RS/signaling overhead

Conclusion

In CSI compression using two-sided model use case, further discuss the pros/cons of different offline training collaboration types including at least the following aspects:

* Whether model can be kept proprietary
* Requirements on privacy-sensitive dataset sharing
* Flexibility to support cell/site/scenario/configuration specific model
* gNB/device specific optimization – i.e., whether hardware-specific optimization of the model is possible, e.g. compilation for the specific hardware
* Model update flexibility after deployment
* feasibility of allowing UE side and NW side to develop/update models separately
* Model performance based on evaluation in 9.2.2.1
* Whether gNB can maintain/store a single/unified model
* Whether UE device can maintain/store a single/unified model
* Extendability: to train new UE-side model compatible with NW-side model in use; Or to train new NW-side model compatible with UE-side model in use
* Whether training data distribution can be matched to the device that will use the model for inference
* Whether device capability can be considered for model development
* Other aspects are not precluded
* Note: training data collection and dataset/model delivery will be discussed separately

Agreement

* In CSI compression using two-sided model use case, further study the necessity, feasibility, and potential specification impact of UE side data collection enhancement including at least
* Enhancement of CSI-RS configuration to enable higher accuracy measurement.
* Assistance information for UE data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc.
  + The provision of assistance information needs to consider feasibility of disclosing proprietary information to the other side.
* Signaling for triggering the data collection
* In CSI compression using two-sided model use case, further discuss the necessity, feasibility, and potential specification impact for NW side data collection including at least:
* Enhancement of SRS and/or CSI-RS measurement and/or CSI reporting to enable higher accuracy measurement.
* Contents of the ground-truth CSI including:
  + Data sample type, e.g., precoding matrix, channel matrix etc.
  + Data sample format: scaler quantization and/or codebook-based quantization (e.g., e-type II like).
  + Assistance information (e.g., time stamps, and/or cell ID, Assistance information for Network data collection for categorizing the data in forms of ID for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc., and data quality indicator)
* Latency requirement for data collection
* Signaling for triggering the data collection

Agreement

In CSI compression using two-sided model use case, further study the following aspects for CSI configuration and report:

* NW configuration to determine CSI payload size, e.g., possible CSI payload size, possible rank restriction and/or other related configuration.
* How UE determines/reports the actual CSI payload size and/or other CSI related information within constraints configured by the network.

Agreement

In CSI compression using two-sided model use case, further study the feasibility and methods to support the legacy CSI reporting principles including at least:

* The priority rule regarding CSI collision handling and CSI omission
* Codebook subset restriction
* CSI processing Unit

Agreement

In CSI compression using two-sided model use case, further study the necessity, feasibility, and potential specification impact for intermediate KPIs based monitoring including at least:

* NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE or obtained from the UE-side.
* UE-side monitoring based on the output of the CSI reconstruction model, subject to the aligned format, associated to the CSI report, indicated by the NW or obtained from the network side.
  + Network may configure a threshold criterion to facilitate UE to perform model monitoring.
* UE-side monitoring based on the output of the CSI reconstruction model at the UE-side
  + Note: CSI reconstruction model at the UE-side can be the same or different comparing to the actual CSI reconstruction model used at the NW-side.
  + Network may configure a threshold criterion to facilitate UE to perform model monitoring.
* FFS: Other solutions, e.g., UE-side uses a model that directly outputs intermediate KPI. Network-side monitoring based on target CSI measured via SRS from the UE.

Note: Monitoring approaches not based on intermediate KPI are not precluded

Note: the study of intermediate KPIs based monitoring should take into account the monitoring reliability (accuracy), overhead, complexity, and latency.

# Reference

R1-2302320 Discussion on other aspects of AI/ML for CSI feedback enhancement FUTUREWEI

R1-2302359 Discussion on AI/ML for CSI feedback enhancement Huawei, HiSilicon

R1-2302438 Discussion on other aspects for AI CSI feedback enhancement ZTE

R1-2302478 Other aspects on AI/ML for CSI feedback enhancement vivo

R1-2302541 On sub use cases and other aspects of AI/ML for CSI feedback enhancement OPPO

R1-2302594 Discussion on other aspects on AIML for CSI feedback Spreadtrum Communications

R1-2302629 Other aspects on ML for CSI feedback enhancement Nokia, Nokia Shanghai Bell

R1-2302696 Discussion on AI/ML-based CSI feedback enhancement CATT

R1-2302750 Discussion on AI/ML for CSI feedback enhancement NEC

R1-2302791 On other aspects on AI/ML for CSI feedback Intel Corporation

R1-2302823 Discussion on AI/ML for CSI feedback enhancement InterDigital, Inc.

R1-2302842 Considerations on CSI measurement enhancements via AI/ML Sony

R1-2302905 Views on specification impact for CSI feedback enhancement Fujitsu

R1-2302919 Discussion on AI-CSI Ericsson

R1-2302976 Discussion on specification impact for CSI feedback based on AI/ML Xiaomi

R1-2303026 Discussion on AI/ML for CSI feedback enhancement China Telecom

R1-2303038 Discussion on AI/ML for CSI feedback enhancement Panasonic

R1-2303051 On Enhancement of AI/ML based CSI Google

R1-2303077 Other aspects on AI/ML for CSI feedback enhancement LG Electronics

R1-2303121 Discussion on potential specification impact for CSI feedback enhancement Samsung

R1-2303184 Discussions on AI-ML for CSI feedback CAICT

R1-2303195 Discussion on other aspects on AI/ML for CSI feedback enhancement ETRI

R1-2303225 Discussion on other aspects on AI/ML for CSI feedback enhancement CMCC

R1-2303337 Other aspects on AI/ML for CSI feedback enhancement MediaTek Inc.

R1-2303436 AI and ML for CSI feedback enhancement NVIDIA

R1-2303476 Discussion on other aspects of AI/ML for CSI enhancement Apple

R1-2303525 Further aspects of AI/ML for CSI feedback Lenovo

R1-2303583 Other aspects on AI/ML for CSI feedback enhancement Qualcomm Incorporated

R1-2303655 Discussion on AI/ML for CSI feedback enhancement AT&T

R1-2303706 Discussion on other aspects on AI/ML for CSI feedback enhancement NTT DOCOMO, INC.

R1-2303810 Discussions on CSI measurement enhancement for AI/ML communication TCL Communication Ltd.