**3GPP TSG-RAN WG3 #113e** ***Draft R3-214220***

**Online, 16th – 26th August 2021**

**Agenda Item: 18.2**

**Source: Lenovo, Motorola Mobility**

**Title: Summary of offline discussion on AI RAN general framework**

**Document for: Discussion and Approval**

# 1 Introduction

**CB: # AIRAN2\_GeneralandFramework**

**- Converge on the open issues on the AI functional framework:**

**Whether Model performance feedback/Model deployment is needed?**

**Data from Model inference?**

**- Performance Evaluation of AI/ML Models?**

**- Update/correct high level principles, definitions and terms, if needed**

**- Provide TPs if agreeable**

(Lenovo - moderator)

Summary of offline disc in [R3-214220](Inbox\R3-214220.zip)

The offline discussion will comprise 2 phases

* Phase 1: Try to identify easy agreements and controversial issues for Phase 2 discussion
  + **Deadline: August 20th, Friday, 4am UTC**
* Phase 2: Try to come up with TP if agreeable
  + **Deadline: August 24th, Tuesday, 4am UTC**

# 2 For the Chairman’s Notes (Phase 1)

# 3 Discussion (Phase 1)

## 3.1 High-Level Principles

In the submitted draft TR R3-214129, based on the agreed TP R3-212978 from RAN3#112e meeting, the following high-level principles are captured:

|  |
| --- |
| **R3-214129 TR 37.817 v0.2.0 (China Mobile Com. Corporation)**  4.1 High-level Principles  The following high level principles should be applied for AI-enabled RAN intelligence:   * The detailed AI/ML algorithms and models for use cases are out of RAN3 scope. * The study focuses on AI/ML functionality and corresponding types of inputs/outputs. * The input/output and the location of Model inference function should be studied case by case. * RAN3 should focus on the analysis of data needed at the Model training function from external functions, while the aspects of how the Model training function uses inputs to train a model are out of RAN3 scope. * Where AI/ML functionality resides within the current RAN architecture, depends on deployment and on the specific use cases. * The Model training and Model inference functions should be able to request, if needed, specific information to be used to train or execute the AI/ML algorithm and to avoid reception of unnecessary information. The nature of such information depends on the use case and on the algorithm. * The Model inference function should signal the outputs of the model only to nodes that have explicitly requested them (e.g. via subscription), or nodes that are subject to actions based on the output from model inference. * NG-RAN is prioritized; EN-DC is included in the scope. FFS on whether MR-DC should be down-prioritized. * A general framework and workflow for AI/ML optimization should be defined and captured in the TR. The generalized workflow should not prevent to “think beyond” the workflow if the use case requires so. |

Besides the agreed principle above, the following principles are proposed by companies to be agreed this time:

1. RAN3 should focus on the analysis of data needed at the Model Inference Function from external functions, while the aspects of how the Model Inference Function uses inputs to derive outputs are out of RAN3 scope [3][5][10]
2. Aspects of how Model Training function performs model deployment/update are out of scope of this study. [3]
3. Aspects of how Model Inference function generates model performance feedback are out of scope of this study. [3]
4. RAN should not act as a data storage or data memory, user data privacy should be respected during AI/ML operation. [12]

**Question 1: Companies are kindly asked if you agree with any/which of the above proposed high-level principles and needs to be captured in the TR in addition?**

|  |  |  |
| --- | --- | --- |
| **Company** | **1), 2), 3), or 4)** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Another issue raised by companies [1][2][3][6], is that an AI/ML model used in a Model Inference function instantiated in a logical RAN node has to be initially trained and tested before deployment (typically in an offline manner with test data). Also, a model used initially in the case of online training in a logical RAN node has to be a trained and tested one.

In moderator’s understanding, the model testing/validation before deployment is reasonable implementation of ML model to assist network optimization. The following question here is whether we need to capture the model test/validation in the high level principle, and if there is any impact on the functional framework description, e.g. if the model test/validation happens in the ML training function or ML inference function, and if we have to distinguish testing/validation data from training data and inference data.

**Question 2a: Companies are kindly asked if ML model test/validation needs to be captured as high-level principle?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Question 2b: Companies are further asked if ML model test/validation shall be reflected in the functional framework?**

* **Yes, please further comment where the ML model test/validation happens, and if we need to define test/validation data in addition to training/inference data.**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

It was left open in the last meeting, whether/how to capture the case that output from one model could be used as input to another model. Some companies [2][3][6][12] want to capture such model chaining in the high-level principle or capture in the functional framework description. In particular, [2][3][6] thinks to reflect such model chaining in the functional framework, Model Inference may provide output to data collection too, such that its output can be further provided to another model as input.

**Question 3a: Companies are kindly asked if “output from one model could be used as input to another model” shall be captured as one high-level principle?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Question 3b: Companies are further asked if model chaining shall be reflected in the functional framework, e.g., a solid/dash line drawn from model inference to data collection providing the output?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2 Functional Framework

|  |
| --- |
| **R3-214129 TR 37.817 v0.2.0 (China Mobile Com. Corporation)** 4.2 Functional Framework *Editor’s Note: Data Preparation aspects may be further refined*    Figure 4.2-1: Functional Framework for RAN Intelligence  This section introduces the common terminologies related to the functional framework for RAN intelligence illustrated in Figure 4.2-1.   * Data Collection is a function that provides input data to Model training and Model inference functions. AI/ML algorithm specific pre-processing of data is not carried out in the Data Collection function.  Examples of input data may include measurements from Ues or different network entities, performance feedback, AI/ML model output.   + Training Data: information needed for the AI/ML model training function.   + Inference Data: information needed as an input for the Model inference function to provide a corresponding output. * Model Training is a function that performs the training of the ML model. The Model training function is also responsible for data preparation (e.g. data pre-processing and cleaning, formatting, and transformation of raw data), if required. * Model Inference is a function that provides AI/ML model inference output (e.g. predictions or decisions). The Model inference function is also responsible for data preparation (e.g. data pre-processing and cleaning, formatting, and transformation of raw data), if required. * Actor is a function that receives the output from the Model inference function and triggers or performs corresponding actions. The Actor may trigger actions directed to other entities or to itself. * Feedback: Information that may be needed to derive training or inference data or performance feedback. |

It has been widely discussed about the metric to evaluate the performance of a ML model, and the outcome of a ML model. Those metrics could be sent together with the Inference Output or as part of the Model Performance Feedback. In particular, the following metrics were listed by different companies:

* Accuracy/Uncertainty [1][5][6][8], e.g., for classification task or for prediction task.
* Variance [1], e.g., for regression task
* Dispersion [1], e.g., for iterative optimization task
* …

Besides, [5][8] also propose to send validity time together with the inference output.

On the other hand, [9][10][11][12] believe whether/how those metrics and validity time are used and sent together with the inference output really depends on the exact AI algorithm used. Thus, [9][10][11][12] suggest to study the definition and usage of metrics and validity time case by case.

In moderator’s observation, it seems common understanding that some kind of metrics (e.g. accuracy, uncertainty, variance etc. ) will be used to evaluate the model performance and model output, however, how exactly it is defined and used in the functional framework(e.g. in which step) depends on the outcome of the use case solution discussion. Therefore, at this stage, moderator would suggest postponing introducing the metrics and validity time in the functional framework until how they are exactly used in each use case becomes clear.

**Question 4: Companies are kindly asked if it’s ok to postpone introducing the metrics (e.g., accuracy, uncertainty, variance etc.) and validity time in the functional framework until how they are exactly used in each use case becomes clear?**

* **Yes**
* **No, please further comment what shall be clearly captured in the functional framework and what is the definition/usage.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

As outcome of the last RAN3 meeting, the model deployment and update from ML Training to ML Inference was left FFS. Among the submitted papers, [2][6][8][9][10] think the Model Deployment and Update shall be kept, while [5] thinks it is only applicable to single vendor environment in practice.

**Question 5: Companies are kindly asked if Model Deployment and Update from ML Training to ML Inference is needed in the functional framework?**

* **Yes, please further comment if we shall limit it to single vendor case.**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Regarding whether ML Inference shall provide model performance feedback to ML training, [2][3][6][7][8][9][10] believe it is needed in one way or another to trigger the ML model retrain/update at the ML training in case the model performance degrades. [2] believe the model performance feedback is applicable to both offline training and online training:

* In case of offline model training the Model Performance Feedback interconnection is only an optional interface and needed only if certain information from Model Inference function is suitable for improvement of the initially trained model. This information could be prediction accuracy or similar statistical data achieved with the model during run time, resulting response time, or processing and memory size/load available/required in Model Inference function.
* For online model training, e.g. using a reinforcement learning approach, information from Model Inference function is fed back to the Model Training function to further improve the model according to adaptation of model-related parameter settings. Such information could be e.g. on prediction accuracy or similar statistical data achieved with the model during run time, on output data drifts, output data quality (granularity/pattern), or output data mismatch.

In [9], traffic load prediction was taken as an example, the model performance could be evaluated by the accuracy or confidence interval by comparing the predicted traffic load with the actual measured traffic load in a certain time period. And in a reasonable implementation, the traffic load prediction will make use of the historical measured traffic load, which means the historical traffic load measurement will be provided to the model inference function as “Inference Data” and by nature model inference function can determine the model performance in a certain time period.

[7] thinks the model performance feedback arrow from Model Inference to Model Training does not really apply to reinforcement learning, thus suggests using a dash line arrow and keeping it optional.

On the other hand, [5][11][12] propose to remove the model performance feedback from ML Inference to ML training due to lack of motivation or justification.

**Question 6: Companies are kindly asked whether to keep the model performance feedback from ML inference to ML training in the functional framework?**

* **Yes, please further comment whether it should be optional, e.g., dash line.**
* **No, please further comment on the reason considering the explanation above.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 3 Conclusion

# Reference

1. R3-213196 AI/ML Model Performance and Reliability Evaluation (Futurewei) discussion
2. R3-213211 Discussion on AI/ML-based functional framework for RAN intelligence (Deutsche Telekom AG) other
3. R3-213294 Proposed TP to TR 37.817 on high-level principles and functional framework (NEC) other
4. R3-213373 AI/ML Architecture (Qualcomm Incorporated, Deutsche Telekom) discussion
5. R3-213418 (TP for SON BL CR for TS 37.817): Framework for RAN intelligence (Ericsson) other
6. R3-213468 High level principle and Functional Framework of AI/ML enabled NG-RAN Network (Intel Corporation) discussion
7. R3-213540 Discussion on framework of AI (CATT) discussion
8. R3-213712 Discussion on Functional Framework and High-Level Principles (Samsung) discussion
9. R3-213723 Remaining issues on AI functional framework (Lenovo, Motorola Mobility) discussion
10. R3-213756 Left issue on AI Functional Framework for RAN Intelligence (ZTE Corporation, China Unicom) other
11. R3-213892 (TP for TR 37.817): Further discussions on the AI/ML Framework (Nokia, Nokia Shanghai Bell) other
12. R3-214078 Further discussions on general principles and frame work (Huawei) other

# Annex