**3GPP TSG-SA5 Meeting #157S5-246263**

**Hyderabad, India, 14-18 October 2024**

**Source: Nokia, ZTE**

**Title: Rel-19 pCR TR 28.858 AIML prediction latency**

**Document for: Approval**

**Agenda Item: 6.19.1**

# 1 Decision/action requested

***The group is asked to discuss and agree on the proposal.***

# 2 References

[1] 3GPP TR 28.858 “Study on Artificial Intelligence / Machine Learning (AI/ML) management Phase 2”.

# 3 Rationale

This pCR is to add two use cases and corresponding requirements for AI/ML inference latency.

# 4 Detailed proposal

|  |
| --- |
| **Start of modification** |

#### 5 AI/ML management use cases and requirements

#### 5.1 ML model training

## 5.1.Y AI/ML prediction latency

5.1.Y.1 Description

AI/ML prediction latency refers to the time it takes for an ML model to process an input and produce an output. This is a critical metric for real-time AI/ML enabled use cases where quick responses are essential. It is important to have the awareness of AI/ML prediction latency as well as to allow consumer to indicate the requirements regarding the AI/ML prediction latency.

5.1.Y.2 Use cases

##### 5.1.Y.2.1 AI/ML prediction latency during ML model training

Depending on the use case, the AI/ML training MnS consumer specifies ML model performance requirements that need to be considered during the training process by the ML training MnS producer. For use cases with stringent time constraints, these requirements may include specific targets for the latency of AI/ML prediction computations.

To meet these latency requirements, the ML training MnS producer can employ various optimization techniques during model training (e.g., quantization, pruning). For example, these techniques may include:

By carefully selecting and applying optimization techniques, the ML training MnS producer can ensure that the trained ML model meets the specified latency requirements, thereby enabling real-time performance in critical use cases.

#### 5.1.Y.3 Potential requirements

REQ-AI/ML\_Train\_Latency-1: The 3GPP management system should have a capability allowing an authorized MnS consumer to provide the accepted AI/ML prediction latency requirements.

|  |
| --- |
| **Start of next modification** |

## 5.5 AI/ML inference

## 5.5.Y AI/ML prediction latency

5.5.Y.1 Description

AI/ML prediction latency refers to the time it takes for an ML model to process an input and produce an output. This is a critical metric for real-time AI/ML enabled use cases where quick responses are essential. It is important to have the awareness of AI/ML prediction latency as well as to allow consumer to indicate the requirements regarding the AI/ML prediction latency.

5.5.Y.2 Use cases

##### 5.5.Y.2.1 AI/ML prediction latency during inference

AI/ML prediction latency is a critical factor in use cases with stringent time constraints. During inference, the latency of AI/ML prediction is influenced by several factors, including the hardware platform executing the inference, techniques such as ML model quantization and pruning, and the overall performance of the ML model. To ensure that the AI/ML prediction latency meets the requirements specified by an authorized MnS consumer, it is essential to monitor the achieved AI/ML prediction latency once the ML model is loaded and activated in the AI/ML inference function and report the metric to the AI/ML inference MnS producer.

#### 5.5.Y.3 Potential requirements

REQ-AI/ML\_INF\_Latency-1: The 3GPP management system should have a capability to report to an authorized MnS consumer the achieved AI/ML prediction latency.

|  |
| --- |
| **End of modifications** |