**3GPP TSG-SA5 Meeting #154 *S5-241785***

Changsha, China, 15 April - 19 April 2024

**Source: ZTE Corporation**

**Title: Rel-19 pCR TR 28.908 add use case, requirements and solution for model complexity**

**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 TS 28.105: "Management and orchestration; Artificial Intelligence/Machine Learning (AI/ML) management".

[2] SP-231722, “Study on AIML management - phase 2”

[3] 3GPP TS 38.843, “Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface”

# 3 Rationale

In release 18, complexity issues for NR use cases including training complexity, inference complexity, and LCM related complexity are widely discussed in RAN 1 [3]. SA5 needs to study the requirements on management for Model complexity in R19.

This contribution is related to WT-5 of the FS\_AIML\_MGT\_Ph2 SID [2], which is to investigate the sustainability aspect of AI/ML.

# 4 Detailed proposal

***Start of First change***

# 5 Use cases

## 5.3 Common management capabilities

## 5.3.x Sustainable AI/ML Operation

5.3.x.1 Description

In release 18, complexity issues for NR use cases including training complexity, inference complexity, and LCM related complexity are widely discussed in RAN 1 [3]. SA5 needs to study the requirements on management for Model complexity in R19.

## 5.3.x.2 Use Cases

5.3.x.2.1 Model Complexity

Model complexity refers to the number of parameters, the depth of the model (especially in neural networks), and the computational complexity of the algorithms, which can be characterized by FLOPS (Floating Point Operations Per Second), MAC (Multiply-Accumulate), and MACC (Model Complexity as MAC Counts).

Model complexity should be a critical consideration for training, deployment, and inference phases for AI/ML LCM.

- For training, more complex models typically require more computational resources, including higher CPU/GPU usage and more memory, which can increase the cost and time required for training.

- For deployment, deploying complex models can be costlier due to the need for more powerful hardware or more instances to handle the same amount of inference load compared to simpler models. More complex models can be harder to update and maintain, especially if the underlying assumptions change or if new types of data need to be accommodated.

- For inference, more complex models generally have longer inference times. This can be critical in applications requiring real-time responses. Besides, complex models may offer higher accuracy, the trade-off in terms of inference speed can limit their applicability in time-sensitive environments.

Based on the analysis above, the model complexity needs to be considered in the AI/ML LCM.

### 5.3.x.3 Potential requirements

**REQ- ModelComplexity -01:** The AI/ML MnS Producer should have a capability to report the information of the ML model complexity.

***End of First change***