**3GPP TSG-SA5 Meeting #158 *S5-247061d1***

Orlando, USA, 18 - 22 November 2024

**Source: NEC, Intel, China Unicom, Telecom Italia , ZTE**

**Title: pCR TR 28.858 add conclusion and recommendations**

**Document for: Approval**

**Agenda Item: 6.19.1**

# 1 Decision/action requested

***The group is asked to review and approve the proposal.***

# 2 References

[1] 3GPP TR 28.858 v0.3.0; Study on Artificial Intelligence / Machine Learning (AI/ML) management Phase 2

# 3 Rationale

*The Study on Artificial Intelligence/Machine Learning (AI/ML) Management Phase 2 [1] is nearing its conclusion. The proposal is to add content for Clause 6, which will include conclusions and recommendations.*

# 4 Detailed proposal

***1st change***

6 Conclusions and recommendations

The present technical report describes the enhanced AI/ML management capabilities and services, based on TS 28.105, for the 3GPP 5GS, including the management and orchestration systems, 5GC, and NG-RAN, where AI/ML features or capabilities are employed. The enhanced management capabilities and services are developed based on a wide range of use cases, along with corresponding potential requirements and possible solutions, as described in Clause 5.

Clause 4 introduces key concepts and relevant terminologies foundational to understanding AI/ML management and orchestration capabilities in the 5GS. It defines terms that extend existing Rel-18 definitions as well as highlighting core principles and concepts that support AI/ML lifecycle management capabilities. Clause 5 then elaborates on the specific management capabilities needed for the AI/ML lifecycle, covering various aspects of ML model training, AI/ML inference emulation, ML model deployment, and AI/ML inference.

Moving into the normative specification development phase, it is recommended to specify AI/ML management capabilities aligned with the relevant use cases, requirements, and solutions for AI/ML lifecycle management steps. These capabilities should address specific functionalities defined by 3GPP, including:

- NG-RAN AIML-based Coverage and Capacity Optimization, and NG-RAN AIML-based Network Slicing defined by RAN3,

- Model delivery/transfer as defined by RAN1/2,

- ML model training and AI/ML inference functions for 5GC as defined by SA2, and

- MDA (Management Data Analytics) as defined by SA5.

Furthermore, detailed capabilities should support:

**ML model training**
- ML-Knowledge-based Transfer Learning
- ML pre-training
- ML Fine-tuning
- ML model training for multiple contexts
- ML training data statistics
- ML model confidence
- Management of Reinforcement Learning
- Sustainable AI/ML for ML training
- ML model Distributed training
- Management of Federated Learning
- ML authentication
- AI/ML prediction latency
- ML explainability

**AI/ML inference emulation**
- ML inference emulation
- ML inference emulation environment selection

**AI/ML deployment**
- Enhance the ML model loading use case
- Managing ML model transfer/delivery

**AI/ML inference**
- Coordination between the ML capabilities
- Sustainable AI/ML for AIML inference
- ML remedial action management
- Managing ML models in use in a live network
- AI/ML prediction latency
- ML explainability

***End of changes***