**3GPP TSG-SA5 Meeting #158 *S5-246362Rev1***

**Orlando, USA, 18 - 22 November 2024**

**Source: NEC, Intel**

**Title: New WID on AI/ML management phase 2**

**Document for: Approval**

**Agenda Item: 6.2.1**

3GPP™ Work Item Description

Information on Work Items can be found at <http://www.3gpp.org/Work-Items>
See also the [3GPP Working Procedures](http://www.3gpp.org/specifications-groups/working-procedures), article 39 and the TSG Working Methods in [3GPP TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm)

Title: AI/ML management - phase 2

Acronym: AIML\_MGT\_Ph2

Unique identifier:

{A number to be provided by MCC at the plenary}

Potential target Release: Rel-19

# 1 Impacts

{For Normative work, identify the anticipated impacts. For a Study, identify the scope of the study}

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Affects: | UICC apps | ME | AN | CN | Others (specify) |
| Yes |  |  | x | x |  |
| No | x | x |  |  |  |
| Don't know |  |  |  |  |  |

# 2 Classification of the Work Item and linked work items

## 2.1 Primary classification

### This work item is a …

|  |  |
| --- | --- |
|  | Study  |
| x | Normative – Stage 1 |
| x | Normative – Stage 2 |
| x | Normative – Stage 3 |
|  | Normative – Other\* |

## 2.2 Parent Work Item

|  |
| --- |
| Parent Work / Study Items  |
| Acronym | Working Group | Unique ID | Title (as in 3GPP Work Plan) |
| FS\_AIML\_MGT\_Ph2 | SA5 | 1020007 | Study on AI/ML management - phase 2  |

### 2.3 Other related Work Items and dependencies

|  |
| --- |
| Other related Work /Study Items (if any) |
| Unique ID | Title | Nature of relationship |
| 940084 | Study on AI (Artificial Intelligence)/ML (Machine Learning) for Air interface | Rel-18 AI/ML in NG-RAN for Aire interface to be managed |
| 940039 | Study on AI/ML management  | Rel-18 SA5 study on AI/ML management |
| 980019 | 5GS support for AI/ML-based services | Rel-18 SA2 work on 5GS support for AI/ML-based services |

**Dependency on non-3GPP (draft) specification:** None

# 3 Justification

It is well recognized that nearly all 3GPP Working Groups, including both RAN and SA WGs, have been actively engaged in advancing the standardization of AI/ML-related features and capabilities. Ongoing support for operations, administration, and management (OAM), led by SA5, plays a critical role in lifecycle management (including ML model training, ML model testing, AI/ML inference emulation, ML model deployment and AI/ML inference management) of these AI/ML-based functionalities. As the 5G capabilities evolve towards more AI/ML techniques being adopted, the Release 18 AI/ML lifecycle management framework and capabilities need to be enhanced to support these advancing AI/ML features.

With the rise of sophisticated 5G use cases utilizing the advanced AI/ML techniques, such as Federated Learning, Reinforcement Learning, online and offline training, and distributed training, there is a strong demand for robust management capabilities. As concluded in the TR 28.858 for the Rel-19 study on AI/ML management phase 2, the recommended enhancements are proposed across the AI/ML lifecycle steps of training, emulation, deployment, and inference, to meet the operational demands of 5GS AI/ML functionalities.

* **ML model training**: The management capabilities for the training step encompasses knowledge-based transfer learning, pre-training, fine-tuning, and techniques like distributed and federated training. Capabilities for managing training data statistics, confidence metrics, and authentication are also essential to support reliable, multi-context applications, and multi-vendor AI/ML features. Energy-aware approaches are also integrated for the training step, to ensure an efficient resource utilization, without compromising model performance.
* **AI/ML inference emulation**: Emulation capabilities facilitate a flexible validation of AI/ML models within a controlled emulation environment before live deployment. By managing the emulation process and emulation environment, models can be rigorously tested while maintaining optimized resource consumption.
* **ML model deployment**: Enhancing processes for model loading, while also leveraging the existing solution to address model transfer/delivery across 5GC and RAN domains will streamline ML model deployment.
* **AI/ML inference**: Enhanced coordination and management of inference operations focus on maintaining low-latency predictions, delivering inference explainability, and supporting energy-efficient remedial actions as needed. This also requires resilient and optimized inference management to support live network reliability and sustainability.

By integrating these capabilities with a focus on energy efficiency, across the training and inference phases, the 5G AI/ML environment will achieve both eco-friendly and operationally effective AI/ML performance.

# 4 Objective

The objectives of AI/ML management phase 2 work item are to specify the management capabilities to support AI/ML functions 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.

To achieve these objectives, the following work tasks are defined:

**WT-1**: Specify the AI/ML management capabilities including use cases, requirements and solutions for the relevant AI/ML lifecycle operational steps based on TR 28.858, including:

 **WT-1.1**: Management capabilities for ML model training:

- ML-Knowledge-based Transfer Learning,
- ML pre-training and 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.

**WT-1.2**: Management capabilities for AI/ML inference emulation:

- ML inference emulation,
- ML inference emulation environment selection.

**WT-1.3**: Management capabilities for ML model deployment:

- Enhancements to ML model loading,

- ML model transfer/delivery (to specifically support RAN1/2 AI/ML functions).

**WT-1.4**: Management capabilities for 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.

NOTE: Solutions for the defined use cases shall leverage the already existing specified capabilities in TS 28.105 where applicable.

# 5 Expected Output and Time scale

***{If this WID covers both stage 2 and stage 3, clearly indicate the different completion dates.}***

|  |
| --- |
| New specifications {One line per specification. Create/delete lines as needed} |
| Type  | TS/TR number | Title | For info at TSG#  | For approval at TSG# | Rapporteur |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

|  |
| --- |
| Impacted existing TS/TR {One line per specification. Create/delete lines as needed} |
| TS/TR No. | Description of change  | Target completion plenary# | Remarks |
| TS 28.105 | Enhancement of AI/ML management |  | This TS covers stages 1, 2 and 3 |
|  |  |  |  |

# 6 Work item Rapporteur(s)

# 7 Work item leadership

SA5

# 8 Aspects that involve other WGs

Collaboration with SA1 on AIML related requirements, SA2 on 5GC AIML, SA3 on AIML security, SA6 on Application enablement layer AIML and RAN WGs for related requirements.

# 9 Supporting Individual Members

|  |
| --- |
| Supporting IM name |
| NEC |
| Intel |
| UScellular |
| CATT |
| China Unicom |
| P.I. Works |
| Telecom Italia |
| ZTE |
| Verizon |