**3GPP TSG-SA5 Meeting #162 *S5-254011***

Goteborg, Sweden, 25 - 29 August 2025

|  |
| --- |
| *CR-Form-v12.3* |
| **CHANGE REQUEST** |
|  |
|  | **28.105** | **CR** | **-** | **rev** | **-** | **Current version:** | **19.1.0** |  |
|  |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME |  | Radio Access Network | **X** | Core Network | **X** |

|  |
| --- |
|  |
| ***Title:***  | Input to draftCR TS 28.105 for correct the description for managing AI/ML based capabilities |
|  |  |
| ***Source to WG:*** | Huawei, NEC |
| ***Source to TSG:*** | SA5 |
|  |  |
| ***Work item code:*** | AIML\_MGT\_Ph2 |  | ***Date:*** | 2025-08-15 |
|  |  |  |  |  |
| ***Category:*** | **F** |  | ***Release:*** | Rel-19 |
|  | *Use one of the following categories:****F*** *(correction)****A*** *(mirror corresponding to a change in an earlier release)****B*** *(addition of feature),* ***C*** *(functional modification of feature)****D*** *(editorial modification)*Detailed explanations of the above categories canbe found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | *Use one of the following releases:Rel-8 (Release 8)Rel-9 (Release 9)Rel-10 (Release 10)Rel-11 (Release 11)…Rel-17 (Release 17)Rel-18 (Release 18)Rel-19 (Release 19) Rel-20 (Release 20)* |
|  |  |
| ***Reason for change:*** | 1. The reference in clause 4.X, 4.Y, 4.Z, 4a.2 are unclear, it is suggested to add the clause reference.

2. For clause 4a.2, the description related to NG-RAN is not aligned with TS 38.300. In TS 38.300, the related statements are as following,*- AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the NG-RAN node;**- AI/ML Model Training and AI/ML Model Inference are both located in the NG-RAN node.**AI/ML Model Training follows the definition of the "ML model training" as specified in clause 3.1 of TS 28.105 [64]. An AI/ML Model needs to be trained, validated and tested before deployment for AI/ML Model Inference.**AI/ML Model Inference follows the definition of the "AI/ML inference" as defined in clause 3.1 of TS 28.105 [64].*RAN use “*NG-RAN node* “, “ML model training” and “AI/ML inference” instead of “gNB” “ML training function” and “AI/ML inference”, therefore,it is proposed to correct the description, change “gNB” to “NG-RAN”, change “ML training function” to “ML model training”, change “AI/ML inference function” to “AI/ML inference”.3. For clause 6.5.4.2.1, 6.5.4.2.2 and 6.5.4.2.3, the description of “An NG-RAN AI/ML-based distributed Network Energy Saving capability may use one or more ML models or AI/ML Inference Functions to derive energy saving recommendations.” “An AI/ML-based distributed Mobility Optimization capability may use one or more ML models or AI/ML Inference Functions to derive handover recommendations.” “An NG-RAN AI/ML-based distributed Load Balancing capability may use one or more ML models or AI/ML Inference Functions to derive load balancing recommendations.” are not correct, these functions belong to RAN, but currently, RAN does not have such descriptions. Therefore, it is proposed to update the statements by referring to the description from RAN specification and removing such descriptions.4. For management of distributed training, it can not be used for RAN AI/ML-based use cases, therefore any descriptions involving "gNB" are incorrect.  |
|  |  |
| ***Summary of change:*** | * Add reference clause in the overview for management of AI/ML capabilities for RAN, 5GC and MDA in clause 4.X, 4.Y and 4.Z.
* Change “gNB” to “NG-RAN”, change “ML training function” to “ML model training”, change “AI/ML inference function” to “AI/ML inference”
* Update the description of clause 6.5.4.
* Remove the words “gNB” in clause 6.2b.2.X5.
 |
|  |  |
| ***Consequences if not approved:*** | It is not clear about which use cases for managing AI/ML capabilities can be supported. |
|  |  |
| ***Clauses affected:*** | 4.X, 4.Y, 4.Z, 4a.2, 6.1, 6.5.4.1, 6.5.4.2, 6.5.4.2.1, 6.5.4.2.2, 6.5.4.2.3, 6.2b.2.X5 |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  | **X** |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  | **X** |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  | **X** |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
| ***This CR's revision history:*** |  |

***Start of next change***

4.X1 Management of AI/ML capabilities for RAN

The management of AI/ML capabilities for the RAN covers scenarios where both the ML model training and AI/ML inference are located in the NG-RAN node, as well as scenarios where the ML model training is located in the management system and the AI/ML inference is located in the NG-RAN node.. In either case, the NG-RAN AI/ML-based feature defined in clause 16.20 of TS 38.300 [16] can be supported.

4.X2 Management of AI/ML capabilities for 5GC

The management of AI/ML capabilities for the 5GC covers scenarios where both the ML model training and AI/ML inference functions are located in the 5GC.in this case, the NWDAF feature defined in clause 6 of TS 23.288 [3] can be supported.

4.X3 Management of AI/ML capabilities for MDA

For MDA, the ML training function can be located either inside or outside the MDAFwhile the AI/ML inference function is located in the MDAF. In this case, the MDA capabilities defined in clause 7.2 of TS 28.104 [2] can be supported.

***Next change***

## 4a.2 AI/ML functionalities management scenarios (relation with managed AI/ML features)

The ML training function and/or AI/ML inference function can be located in the RAN domain MnS consumer (e.g. cross-domain management system),a domain-specific management system (i.e. a management function for RAN or CN), or in a network function (NF).

For MDA, the ML training function can be located inside or outside the MDAFwhile the AI/ML inference function is located in the MDAF.

For NWDAF, the ML training function can be located in the MTLF of the NWDAF or in the management system, and the AI/ML inference function is located in the AnLF.

For NG-RAN, the ML training function and AI/ML inference function can both be located in the NG-RAN node.or the ML training function can be located in the management system and AI/ML inference function is located in the NG-RAN node. Where the ML training function corresponds to ML model training that stated in clause 16.20.2 in TS 38.300[16] and AI/ML inference function can correspond to AI/ML inference stated in clause 16.20.2 in TS 38.300[16].

For LMF-based AI/ML Positioning, the ML training function can be located in the LMF or CN-domain management function, and the AI/ML inference function can be located in the LMF.

Therefore, multiple location scenarios for ML training function and AI/ML inference functions are possible.

**Scenario 1:**

The ML training function and AI/ML inference function are both located in the 3GPP management system (e.g. a RAN domain management function). For example, for RAN domain-specific MDA, both the ML training function and AI/ML inference functions for MDA can be located in the RAN domain-specific MDAF as depicted in figure 4a.2-1.



Figure 4a.2-1: Management for RAN domain specific MDAF

Similarly, for CN domain-specific MDA the ML training function and AI/ML inference function can be located in CN domain-specific MDAF.

**Scenario 2:**

For AI/ML for NG-RAN, the ML model training is located in the 3GPP RAN domain-specific management function while the AI/ML inference is located in NG-RAN node. For AI/ML inference use case, refer to Network Energy Saving, Load Balancing, Mobility Optimization as defined in clause 16.20 in TS 38.300. See Figure 4a.2-2.



Figure 4a.2-2: Management where the ML model training is located in RAN domain management function and AI/ML inference is located in NG-RAN node

**Scenario 3:**

For AI/ML in NG-RAN, the ML model training and AI/ML inference are both located in the NG-RAN node. For ML model training and AI/ML inference use case, refer to Network Energy Saving, Load Balancing, Mobility Optimization as defined in clause 16.20 in TS 38.300. See Figure 4a.2-3.



Figure 4a.2-3: Management where the ML model training and AI/ML inference are both located in NG-RAN node

**Scenario 4:**

For NWDAF, both the MTLF and AnLF are located in the NWDAF. See Figure 4a.2-4.



Figure 4a.2-4: Management where the MTLF and AnLF are both located in CN

***Next change***

## 6.1 ML model lifecycle management capabilities

Each operational step in the ML model lifecycle (see clause 4a.0.1) is supported by one or more AI/ML management capabilities, which enable the MnS consumer (e.g. operator) to manage and control the ML model lifecycle as listed below.

**Management capabilities for ML model training**

**- ML model training management**: allowing the MnS consumer to request the ML model training, consume and control the producer-initiated training, and manage the ML model training/re-training process. The training management capability may include training performance management and setting a policy for the producer-initiated ML model training.

**-** ML model training capability also includes validation to evaluate the performance of the ML model when performing on the validation data, and to identify the variance of the performance on the training and validation data. If the variance is not acceptable, the ML model would need to be re-trained before being made available for the next step in the ML model lifecycle (e.g., ML model testing).

**Management capabilities for ML testing**

**- ML model testing management**: allowing the MnS consumer to request the ML model testing, and to receive the testing results for a trained ML model. It may also include capabilities for selecting the specific performance metrics to be used or reported by the ML testing function. MnS consumer may also be allowed to trigger ML model re-training based on the ML model testing performance results.

**Management capabilities for AI/ML inference emulation:**

* **AI/ML inference emulation:** a capability allowing an MnS consumer to request an ML inference emulation for a specific ML model or models (after the training, validation, and testing) to evaluate the inference performance in an emulation environment prior to applying it to the target network or system.

**Management capabilities for ML model** **deployment:**

**- ML model loading management**: allowing the MnS consumer to trigger, control and/or monitor the ML model loading process.

**Management capabilities for AI/ML inference:**

**- AI/ML inference management:** allowing an MnS consumer to control the inference, i.e., activate/deactivate the inference function and/or ML model/models, configure the allowed ranges of the inference output parameters. The capabilities also allow the MnS consumer to monitor and evaluate the inference performance and when needed trigger an update of an ML model or an AI/ML inference function.

The use cases and corresponding requirements for AI/ML management capabilities are specified in the following clauses.

***Next change***

### 6.5.4 AI/ML inference capability configuration management

#### 6.5.4.1 Description

The objective of AI/ML for NG-RAN is to improve network performance and user experience, which can yield further insights, e.g., for Network Energy Saving, Load Balancing, Mobility Optimization as defined in TS 38.300 [16]. According to the principles defined in clause 16.20.2, clause 15.4, clause 15.5 in TS 38.300 [16], either “AI/ML model training is located in the OAM and AI/ML model inference is located in the NG-RAN node;” or “AI/ML model training and AI/ML model inference are both located in the NG-RAN node”. To manage the network performance,the AI/ML inference and the associated ML model(s) may need to be managed and configured to conduct inference in the 5G system in alignement with the consumer´s expectation, e.g., to enable the AI/ML inference function to perform inference.

The MnS producer for AI/ML inference management needs to provide a capability for configuration of the AI/ML inference function.

#### 6.5.4.2 Use cases

##### 6.5.4.2.1 Managing NG-RAN AI/ML-based distributed Network Energy Saving

The MnS consumer monitors the network performance and determines whether, and when, to activate or deactivate an AI/ML inference related to an AI/ML-based Distributed Network Energy Saving. The activation and deactivation actions for AI/ML inference related to an AI/ML-based Distributed Network Energy Saving perfromed by the MnS producer may also be triggered by policies provided by the consumer.

##### 6.5.4.2.2 Managing NG-RAN AI/ML-based distributed Mobility Optimization

The MnS consumer monitors the network performance and determines whether activation or deactivation of an AI/ML inference related to an AI/ML-based Distributed Mobility Optimization is required. The activation and deactivation actions performed by the MnS producer may also be triggered by some defined policies provided by the consumer.

##### 6.5.4.2.3 Managing NG-RAN AI/ML-based distributed Load Balancing

The MnS consumer monitors the network performance and determines whether, and when, to activate or deactivate an AI/ML-based Distributed Load balancing. The activation and deactivation actionsperfrormed by the MnS producer may also be triggered by policies provided by the consumer.

***Next change***

#### 6.2b.2.X5 Management of Distributed Training

Distributed training is a model training approach that involves distributing the training workload across multiple training functions to accelerate the training process and/or reduce the required computational resources.

In 5GS, the ML training function may be located within the management system or in the NF (e.g. NWDAF),. Each training node has different computing resources and storage capacity based on physical infrastructure such as CPU/GPU/DPU, memory, storage, and network bandwidth. In order to obtain load balance between nodes and maximize the efficiency of resource utilization, splitting up the training may be necessary and involving multiple training functions according to the actual situation of nodes may be needed. Thus, aspects of distributed training need to be supported in the management systems.

Distributed training refers to the approach of distributed computation to scale out a training job, either to accelerate the process or to handle workloads that cannot fit into a single machine.

Management of Distributed Training can be used for AI/ML-based use cases specified in [2] and [3].In 5GS, distributed training can apply across various deployment scenarios for the ML training function. These functions may be located within the 3GPP management system, domain-specific management functions (e.g., RAN domain management function or CN domain management function), or directly in Network Functions such as the NWDAF. The location of these functions depends on the specific scenario defined in clause 4a.2.

When receiving an ML training request, the MLT MnS producer may evaluate whether distributed training is needed according to the training requirements provided by the ML training consumer, and it is up to the MLT MnS producer to determine, based on some information (e.g target inference location) provided by the consumer, appropriate training function(s) which may need to participate in the ML model training. The training requirement may further include (not limited to) expected model performance. Collaboration, mutual agreement and authentication procedures are needed to be establish between distributed ML training functions before sharing any information between these functions.

The actions of ML model distributed training may involve for example, splitting the training of an ML model across many ML training functions, each responsible for computing a portion of the ML models operations. Since the training data may be sparse, MLT MnS consumer may provide indication that the training data should not be split while splitting the training among multiple training functions.

NOTE 1: How to split the ML model and synchronize the parameters in different training function depends on the distributed algorithm which are proprietary and not in scope for standardization.

NOTE 2: The data exchange between different training functions should be in the security tunnel with appropriate authentication and authorization mechanisms.

***End of change***