3GPP SA WG5:
AI/ML Management

Hassan Al-Kanani (NEC), Yizhi Yao (Intel)

Joint Workshop 3GPP SA5 & ETSI ZSM
3 July 2025

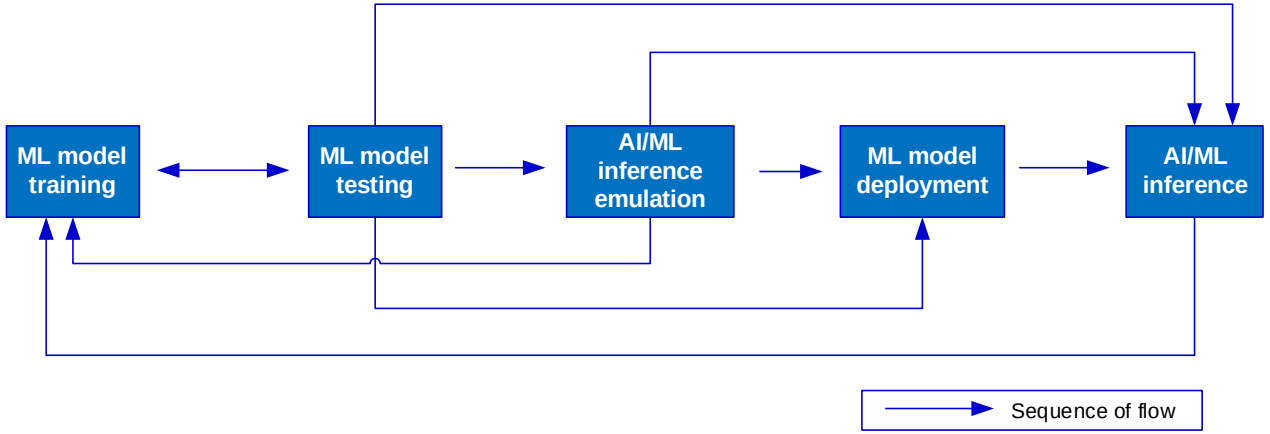
Contents

- **AI/ML in 5GS**
  - Capabilities & Lifecycle Management
  - Lifecycle Management Framework
  - Management Capabilities
- **ML Model Lifecycle Phases**
  - ML model training
  - ML model testing
  - AIML inference emulation
  - ML model deployment
  - AI/ML Inference
- **Deployment Scenarios**
- **3GPP Progress & planned work**
  - Rel-18/19 Accomplishments
  - Rel-20 (5GA) Planned Work
- **References**

AI/ML Capabilities & Lifecycle Management in 5GS

- AI/ML capabilities, such as Management Data Analytics (MDA), Network Data Analytics Function (NWDAF), and RAN intelligent functions, are increasingly integral to the 5G System (5GS), providing advanced automation, prediction, and optimization features that are actively addressed by relevant 3GPP Working Groups.
- These capabilities are powered by ML models, which are executed via AI/ML inference functions deployed across the system.
- To ensure operational effectiveness, adaptability, and scalability across diverse use cases, comprehensive management of ML models and inference functions is essential.
- This management spans key lifecycle phases:
  - **ML Model training:** Generation and update of models using relevant data.
  - **ML Model testing:** Evaluation of model performance based on testing data before deployment.
  - **AI/ML inference emulation:** Pre-deployment emulation of model inference to assess behaviour and outputs.
  - **ML Model deployment:** Distribution of the trained model in the system.
  - **AI/ML inference:** Real-time or scheduled execution of the model to deliver intelligent functionality.

AI/ML lifecycle management framework



ML Model Training:

- Initial training & re-training.
- Validates model performance using validation data (triggers re-training if validation results are unsatisfactory).

ML Model Testing:

- Evaluates the trained ML model using testing data.
- If performance criteria are not met, retraining is triggered.

AI/ML Inference Emulation (Optional):

- Emulate inference execution in a non-production environment to assess performance before deployment.
- Optional step, used to mitigate risks prior to deployment.
- If results do not meet expectations (e.g., performance issues or impact on existing functionalities), re-training is required.

ML Model Deployment:

- Loads the trained ML model on to the target AI/ML inference function.
- May be skipped if training and inference functions are co-located.

AI/ML Inference:

- Performs inference using the trained ML model.
- May trigger model re-training or updates based on performance evaluation.

Note: *A comprehensive set of corresponding terminologies has been developed to support understanding and to guide the development of management capabilities within 5GS and beyond.*

Management Capabilities

- **Management Capabilities for ML Model training**
  - **ML Model training management**: Enables MnS consumer to request, monitor, and control ML model training and re-training, policy settings for producer-initiated training.
  - **Training Performance Management**: Evaluate how well ML models meet training objectives, select and use performance indicators to monitor and assess training effectiveness.
  - **Validation management**: Evaluates ML model performance using validation data; triggers re-training if performance variance is unacceptable.
- **Management Capabilities for ML testing**
  - **ML Model testing management**: Allows MnS consumer to request testing and receive performance results.
  - **Performance metrics selection**: Enables choice of specific metrics for ML testing.
  - **Re-training triggers**: Allows MnS consumer to initiate re-training based on test results.

- **Management Capabilities for AI/ML inference emulation**
  - **AI/ML Inference emulation management**: Enables MnS consumer to request inference emulation and receive emulation report for performance evaluation before deployment.
- **Management Capabilities for ML Model deployment**
  - **ML Model loading management**: Allows MnS consumer to trigger, control, and monitor the ML model loading process.
- **Management Capabilities for AI/ML inference**
  - **AI/ML inference management**:
    - Activates/deactivates inference function or specific ML models.
    - Configures allowed inference output parameters.
    - Monitors and evaluates inference performance.
    - Triggers updates for ML models or AI/ML inference functions as needed.

ML model training

- Includes both training types, i.e., **initial training**: model is trained from scratch using training data, and **re-training:** performed to improve or restore model performance when degradation is detected or new data becomes available without changing the model structure.
- Use cases:
  - **Consumer-requested training**: Consumer requests model training for specific inference needs.
  - **Producer-initiated training**: Triggered automatically if performance degrades or new data appears.
  - **Model selection**: Consumers choose models suited to local conditions (e.g., urban/rural).
  - **Training process control**: Start, suspend, resume training tasks as needed.
  - **Error handling:** Assess input data quality and manage confidence scores.
  - **Joint training**: Multiple models trained together for coordinated inference.
  - **Training data effectiveness**: Evaluate and report data sample contributions.

ML model training – Performance management

- Defines how ML model training is assessed to ensure performance goals are met using relevant indicators. Supports multi-model, multi-metric evaluation and flexible selection aligned with use cases and consumer policies.
- Use Cases:
  - **Multi-Model, Multi-Metric Evaluation**: Multiple algorithms can be evaluated together, each with one or more performance indicators (e.g., accuracy, F1 score, MSE).
  - **Indicator Query & Custom Selection**: Consumers query supported indicators, then select a subset based on their specific use case and goals.
  - **Policy-Based Indicator Selection**: Consumers may define high-level goals (e.g., “high reliability”); producers translate these into suitable technical metrics — helpful when consumers have limited technical expertise.
- **Performance Indicators**:
  - Accuracy, Precision, Recall, F1 Score, MSE, MAE, RMSE

ML model testing

- Verifies that trained models meet performance requirements using testing data before deployment; testing can be initiated by either party.
- Use Cases:
  - **Consumer-requested testing**: Consumers request performance checks after training.
  - **Producer-initiated testing**: Triggered automatically by the producer after training or validation.
  - **Joint testing**: Multiple models tested together for coordinated deployment readiness.

ML model testing – Performance management

- Evaluate ML model performance during the testing phase using predefined indicators and flexible selection. Supports consumer choice, producer reporting, and policy-based configurations.
- Use Cases:
  - **Multi-Model Testing with Performance Indicators**: Allows testing of one or more models together, using multiple performance metrics such as Accuracy, Precision, Recall, F1 Score, MSE, MAE, RMSE.
  - **Consumer Selection & Producer Support**: Consumers select preferred indicators for evaluation, while the MnS producer provides the list of supported metrics and generates results.
  - **Policy-Based Performance Configuration**: Consumers can define high-level performance expectations through policies, which the producer translates into detailed indicator reporting for decision-making.
- **Performance Indicators**:
  - Accuracy, Precision, Recall, F1 Score, MSE, MAE, RMSE

AI/ML inference emulation

- Inference emulation, performed before production deployment, allows MnS consumers to validate that ML models or inference functions behave correctly and satisfactorily under expected runtime conditions.
- Use Case:
  - **Inference emulation request**: Consumers request emulation to assess behaviour and performance under expected conditions.

ML model deployment

- Manages the distribution of trained models after successful training, validation, and optional emulation.
- Use Cases:
  - **Consumer-requested loading:** Consumers initiate loading and monitor activation progress.
  - **Producer-initiated loading:** Producers load models autonomously based on policies.
  - **Model registration**: Tracks versions and metadata for retrieval and reproducibility.

AI/ML inference – Performance management

- Executes trained models to deliver outputs that enable network automation, with continuous monitoring for accuracy.
- Use Cases:
  - **Runtime monitoring:** Assess inference behaviour to detect performance issues.
  - **Performance evaluation**: Evaluate outputs and their impact on network KPIs.
  - **Event-triggered assessments**: Run evaluations periodically or when conditions change.

AI/ML inference - Update control & History tracking

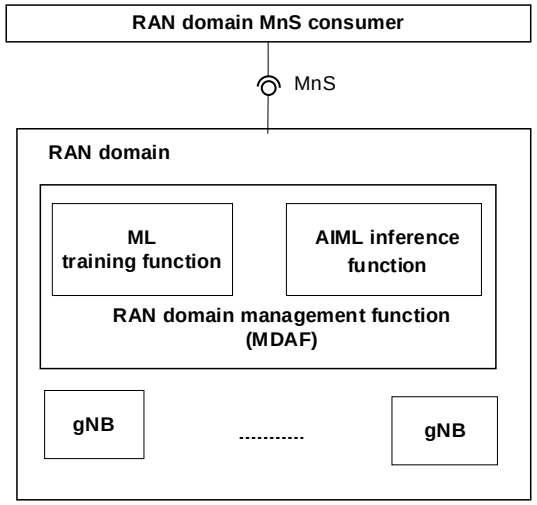
- As network conditions evolve, the performance or suitability of deployed ML models may degrade. AI/ML update control allows MnS consumers to request updated models or capabilities from inference functions, ensuring continued relevance and accuracy. Updates may involve loading new models or triggering full retraining workflows.
- To improve visibility, accountability, and future optimisation, it is important to retain a history of AI/ML inference outcomes along with the contextual conditions under which those inferences were made. This historical insight supports traceability, diagnostics, and adaptive learning.
- Use cases
  - **Availability of new capabilities or ML models**
    - AI/ML inference functions may self-update or learn over time.
    - Consumers may request to be notified when enhanced capabilities or updated ML models become available.
    - Consumers can request updates when inference performance degrades.
  - **Triggering Model Updates**
    - Consumers can request model updates when inference performance degrades.
    - The producer may act by retraining, loading a new model, or triggering remote update processes.
    - Update progress and outcomes can be reported back to the consumer.
    - Policies may be set by the consumer to manage update frequency, timing, and performance targets.
  - **Tracking Inferences and Context**
    - Tracking Inferences and Context
    - Records inference outputs along with relevant context (e.g., network load, time, weather).
    - Enables consumers to analyse inference appropriateness, trends, or model degradation.
    - Supports reporting and control of inference history compilation via the MnS producer.

AI/ML inference – Capabilities & configuration management

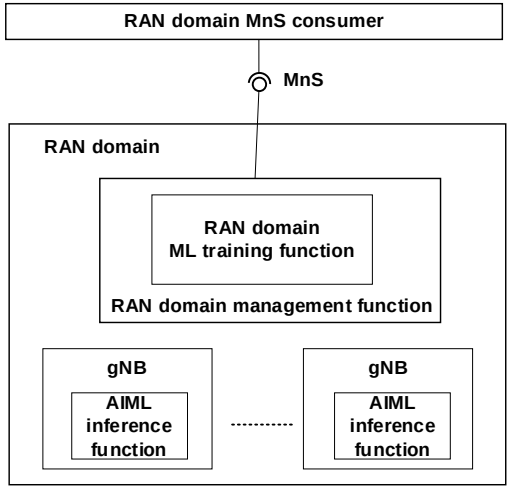
- Effective use of AI/ML in the network relies on understanding what each model can do and configuring its behaviour to adapt to network needs.
- Capabilities management enables consumers to discover and align ML model functions with their automation or intent-based needs, while configuration management supports activation, deactivation, and policy-driven control of inference functions for use-case-specific optimisation across NG-RAN functions.
- Use cases
  - **Identifying ML Model Capabilities:**
    - Consumers request details about available inference capabilities.
    - Supports automation by matching available model functions to operational goals.
  - **Mapping Capabilities to Outcomes**
    - Enables mapping between desired outcomes (e.g., intent fulfilment) and suitable ML models.
    - Supports orchestration of one or more models, including complex, multi-model workflows.
  - **Configuration management:**
    - Adjust inference for energy saving, mobility optimisation, load balancing.

AI/ML functionalities management scenarios

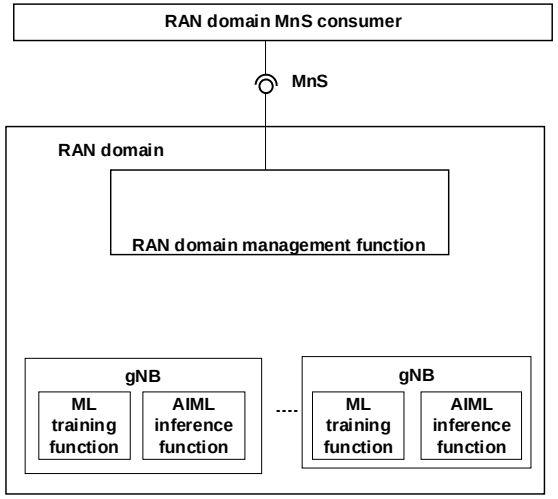
- The **ML training** and **AI/ML inference functions** can reside in the same or different parts of the 5G system depending on the deployment scenario:
  - RAN or CN Management Systems,
  - Network Functions (e.g., gNB, NWDAF),
  - Cross-domain Management Systems.



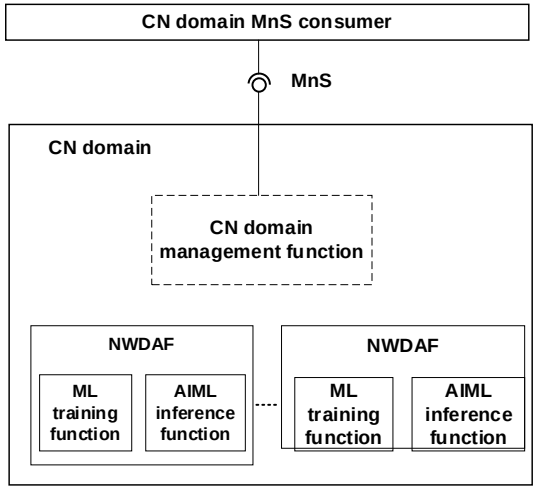
Scenario 1: Both in Management System - Both ML training and inference functions reside in the RAN or CN domain-specific management function (e.g., MDAF).



Scenario 2: Split between Management and gNB - ML training is in the RAN management function; inference runs in the gNB.

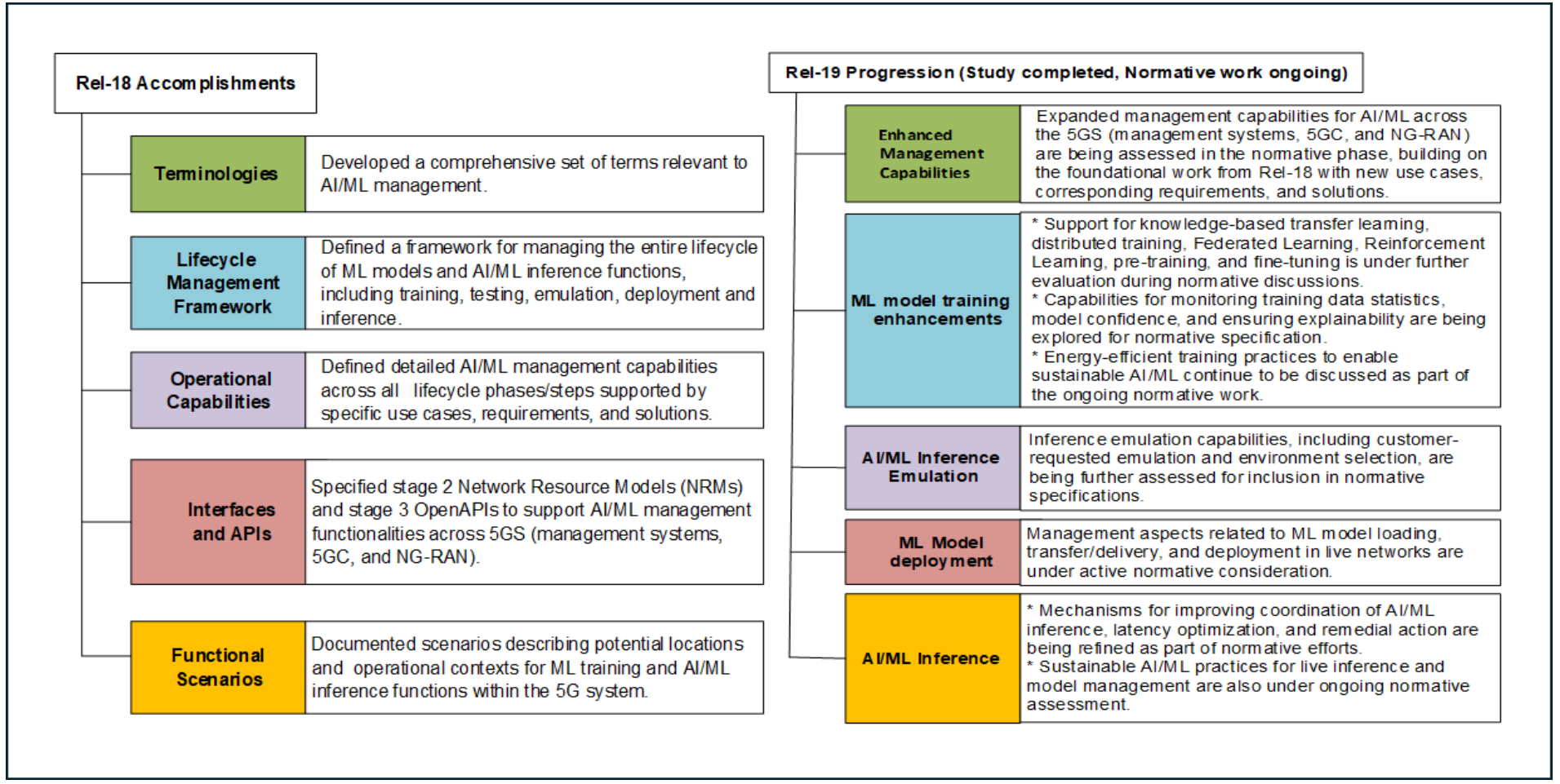


Scenario 3: Both in gNB - Both ML training and inference occur directly in the gNB, enabling localized intelligence.



Scenario 4: Both in NWDAF - For Core Network domain, both functions are hosted in NWDAF (training in MTLF, inference in AnLF).

Rel-18/19 accomplishments



Rel-20 (5GA) – Planned work

WT-1: AI/ML Lifecycle Management Enhancements

WT-1.1: Investigate enhancements of AI/ML management capabilities throughout the AI/ML lifecycle in 5GS, including training, testing, emulation, deployment, inference, to support AI/ML-enabled features in the 5GS. This includes:

1. ML model transfer/delivery as defined by RAN for Solution 4b: OAM can transfer/delivery AI/ML model(s) to UE.
2. NG-RAN use cases including QoE optimization, network energy saving, and mobility use case(as defined in RP-250812).
3. 5GC Analytics: Encompasses new 5GC analytics use cases currently under study under WT#2 (see SP-250413) and investigates OAM support for provisioning ML models to relevant 5GC functions to enable AI/ML-based analytics.
4. LMF-based AI/ML Positioning including data collection and ML model training by the OAM for UE positioning.
5. Study feasibility and potential requirements for data collection for (e.g., UE-side and Network-side) to enable model training.

NOTE 1: The works in the subtasks 1-5 listed above is subject to the progress in the relevant WGs in RAN and SA.

WT-1.2: Investigate enhancements of management capabilities to address the specific needs of selected AI/ML training and inference technologies relevant to 5GS. The focus will include Federated Learning, Distributed learning, Reinforcement Learning and Fine-tuning, which are applicable across AI/ML-based functionalities in RAN, 5GC, and OAM.

WT-2: AI/ML Sustainability

WT-2.1: Investigate the development of specific metrics and evaluation methods to assess and optimize the energy consumption, efficiency, and resource utilization of ML models and AI/ML inference functions across the relevant lifecycle steps.

WT-2.3: Investigate management enhancements to enable monitoring, reporting, and management related to the use of renewable energy sources (e.g., solar, wind, hydro) in AI/ML model training and inference operations.

NOTE 2: Where applicable, alignment with existing energy efficiency management principles as specified in TS 28.310 will be considered.

WT-3: Relation with other management capabilities

WT-3.1: Investigate relation between AI/ML and other management capabilities (e.g. data management).

WT-4: Registration and Discovery management for AI/ML

WT-4.1: Study enhancements to support registration and discovery management for AI/ML

NOTE 3: The work for WT-4.1 will ensure alignment with existing Management and Orchestration (MnS) discovery mechanisms.

References

- 3GPP Rel-18 TS 28.105; Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management.
- 3GPP Rel-19 TS 28.105; Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management.
- 3GPP Rel-18 TR 28.908; Study on Artificial Intelligence/Machine Learning (AI/ ML) management
- 3GPP Rel-19 TR 28.858; Study on Artificial Intelligence / Machine Learning (AI/ML) management phase 2.

Thank You !