ETSI GR ENI 045 V4.1.1 (2025-10)

GROUP REPORT

Experiential Networked Intelligence (ENI); Research on Application Scenarios of Network Large Language Models for Operation, Administration, Maintenance, and Performance

Disclaimer

The present document has been produced and approved by the Experiential Networked Intelligence (ENI) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/ENI-0045v411_OAM LL

Keywords
administration, large language model,
maintenance, operation, performance

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Experiential Networked Intelligence (ENI).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document focuses on how to leverage large language model technologies to assist in communication network operations and management. It identifies typical application scenarios, analyses key technologies, and investigates the business architecture and standardization requirements for applying large language model technologies in communication network operations and management scenarios. Specific technical aspects include:

1) Analysis of how large language models can assist communication network operations and management.
2) Application scenarios of communication network operations and management assisted by large language models.
3) Interaction processes in communication network operations and management assisted by large language models.
4) Standardization recommendation for communication network operations and management assisted by large language models.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] ETSI GR ENI 004: "Experiential Networked Intelligence (ENI); Terminology".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in ETSI GR ENI 004 [i.1] and the following apply:

AI-Enhanced Network OAMP System: application or platform providing network OAMP capabilities by integrating the Network OAMP LLM Service with relevant data, tools, and workflows

NOTE: This represents the complete system presented to end-users or other integrated systems. It combines the AI capabilities of the Network OAMP LLM Service with components like user interfaces, data connectors, reporting tools, automation scripts, and potentially traditional OAMP functions.

Hallucination: statement that the most statistically probable sequence of words does not correspond to factual reality

Hybrid AI Architecture: system design that strategically integrates and orchestrates a portfolio of different Artificial Intelligence models and analytical tools, rather than relying on a single, monolithic model to handle all tasks

NOTE: In the context of OAMP, this architecture is managed by the AI-Enhanced Network OAMP System, which acts as a central orchestrator.

Network OAMP LLM: large language model trained on network Operation, Administration, Maintenance, and Performance data

NOTE: A network OAMP LLM is typically created by fine-tuning a foundation LLM (either general-purpose or ideally domain-specific) with specialized data and tasks relevant to network operation, administration, maintenance, and performance.

Network OAMP LLM Service: deployed service providing programmatic access to the Network OAMP LLM

NOTE: This typically includes the inference endpoint, runtime environment, version management, and necessary wrappers for the Network OAMP LLM, enabling other systems to utilize its capabilities via an interface (e.g. an API).

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

5G	Fifth Generation
AI	Artificial Intelligence
API	Application Programming Interface
HITL	Human-In-The-Loop
ICT	Information and Communications Technology
LLM	Large Language Model
ML	Machine Learning
MLOp	Machine Learning Operations
NE	Network Entity
NF	Network Function
NMS	Network Management System
NOC	Network Operations Center
OAMP	Operation, Administration, Maintenance, and Performance
RAG	Retrieval Augmented Generation
RAN	Radio Access Network
RCA	Root Cause Analysis
SDO	Standards Development Organization
SPN	Slicing Packet Network
SQL	Structured Query Language
TMN	Telecommunications Management Network
VM	Virtual Machine

4 Introduction

Large-scale pre-trained models (referred to as large language models or LLMs) learn patterns statistically to model text from a vast amount of labelled and unlabelled data. The parameters of LLMs encode patterns and relationships learned from the training data. These patterns allow the model to generate responses to queries. However, this capability is a sophisticated form of statistical pattern matching, not the storage of explicit facts or the use of cognition to reason.

By encoding patterns and relationships from training data into a large number of parameters, and then fine-tuning these parameters for specific tasks, the LLM leverages their learned representations to perform well on various downstream tasks. Adopting pre-trained foundation models, such as LLMs, and fine-tuning them for specific downstream tasks, rather than learning models from scratch, has become a new paradigm in AI/ML applications. This approach leverages the broad knowledge captured in these models to improve performance on specialized tasks. In the field of network OAMP, the use of such fine-tuned LLMs is increasingly being explored for large-scale applications.

The high operational costs of training and inference for LLMs necessitate a collaborative approach between LLM providers, LLM consumers, and application integrators. Standards for AI and ML are starting to gain traction (e.g. JTC1/SC42) as well as work within Standards Development Organizations (SDOs) like 3GPP, there are currently no standards that address network Operation, Administration and Maintenance in the context of LLMs.

NOTE: Although it possibly seems counter intuitive for LLM providers to share costs, the telecommunications sector is a potentially lucrative market. Collaboration in this field could provide both long-term relationships and revenue streams, as well as valuable industry-specific insights for LLM providers.

However, in order to be used in mission-critical environments, shortcomings of LLMs need to be addressed. For example, LLMs probabilistically predict the next word, token, etc. gives rise to "hallucinations." This is because responses are constructed based on their statistical likelihood to appear together, not based on facts. A hallucination occurs when the most statistically probable sequence of words does not correspond to factual reality. Put another way, the model is not reasoning or verifying information; it is simply completing a sophisticated pattern, and that pattern results in statements that are coherent yet entirely fabricated.

This research focuses on how to leverage large language model technologies to enhance telecommunication network operations and management. Specifically, it will:

1) Identify typical application scenarios where LLMs are able to enhance network OAMP operations.

2) Analyse key technologies, including fine-tuning methods and integration strategies as well as hallucination mitigation methods.

3) Investigate the service architecture and potential return on investment for applying large language model technologies in telecommunication network OAMP scenarios.

5 LLM-Assisted Network OAMP Operations

5.1 Introduction

This clause defines the primary roles, actors, and architectural components of a system that uses a Large Language Model (LLM) to assist with network Operation, Administration, Maintenance, and Performance (OAMP) tasks.

5.2 Roles and Actors

A "Role" defines a set of responsibilities and permissions, while an "Actor" is the person, group, or system performing that role.

Roles and actors in network OAMP LLM system include:

a) **End-User Role**:
This represents the primary individuals who interact with the AI-Enhanced Network OAMP System to perform or receive assistance with network OAMP tasks (e.g. troubleshooting faults, analysing performance data, planning maintenance). This role focuses on using the system's capabilities to achieve operational objectives. Typical Actors include Network Operators, NOC Technicians, Performance Analysts, Field Engineers, and Network Planners.

b) **AI Maintainer Role**:
This role is responsible for the lifecycle management and operational health of the Network OAMP LLM Service and its underlying LLM. This role focuses on the technical management of the AI components. Key activities include deploying model updates, monitoring the AI service's performance and resource usage, troubleshooting AI-specific issues, managing API access (if applicable), and potentially coordinating model retraining or fine-tuning activities. Typical Actors include MLOps Engineers, AI/ML Engineers, specialized Platform Administrators, and Data Scientists.

5.3 Architectural Framework

The framework for an AI-Enhanced Network OAMP System consists of the following:

a) **AI-Enhanced Network OAMP System**:
The primary application or platform that end-users interact with. This system integrates the Network OAMP LLM Service with relevant, and often siloed, network data sources, operational tools, and business workflows. It serves as the orchestrator, managing the logic for when and how to call the LLM Service, process its responses, and interact with other network systems to fulfil an OAMP task.

NOTE 1: While its scope covers traditional OAMP functional areas (inspired by frameworks like ITU-T TMN), its defining characteristic is the integration of AI to enhance these functions.

b) **Network OAMP LLM Service**:
This is a deployed and managed service that provides programmatic access (e.g. via API) to the Network OAMP LLM. It is responsible for handling inference requests from authorized users (e.g. the AI-Enhanced Network OAMP System), managing the model's runtime environment (including scaling and versioning), and returning the LLM's generated responses to authorized users.

NOTE 2: This service encapsulates the specialized MLOps aspects of the AI model, abstracting the complexity of running the LLM from the primary OAMP application.

c) **Network OAMP LLM**:
This is an LLM that has been trained or fine-tuned on network OAMP data and tasks.

NOTE 3: This model forms the core AI engine, specialized for reasoning and generation tasks within the network OAMP domain.

d) **Support components and techniques**:
This is a set of essential elements and methods used to develop, enhance, and evaluate the System. These are often integrated within the AI-Enhanced Network OAMP System or MLOps workflows and include:

a. Retrieval-Augmented Generation (RAG): a technique used to improve the factual grounding and relevance of LLM responses. It works by retrieving relevant information from external knowledge sources (e.g. technical manuals, real-time network databases, trouble tickets) and providing it to the LLM as context along with the user's query.

b. Prompt Engineering Resources: a curated collection of optimized prompt templates and guidelines. These resources are used to structure queries to the Network OAMP LLM to elicit the most accurate and relevant responses for specific tasks.

c. Evaluation Frameworks: The tools and methodologies used to systematically assess the performance, accuracy, and safety of the Network OAMP LLM's responses. This is critical for both initial development and ongoing monitoring to measure factual accuracy and relevance, and to detect issues like toxicity or bias

6 Application scenarios in large language model assisted network OAMP

6.1 Overview

Network OAMP encompasses a wide range of application scenarios, including single-domain use cases for networks such as RAN, core and transmission networks, as well as cross-domain OAMP use cases such as cross-domain troubleshooting or performance assurance. LLMs are applied to enhance these scenarios by acting as sophisticated reasoning and language processing engines. The primary categories of LLM application in network OAMP are:

- Knowledge-based Q&A for network OAMP: The LLM's ability to understand natural language queries and generate coherent, context-aware responses enables the creation of powerful Q&A systems for OAMP staff. The LLM synthesizes answers by leveraging knowledge from its training data and by reasoning over factual, up-to-date information retrieved from external knowledge bases (e.g. technical manuals, network topology databases) using techniques like Retrieval-Augmented Generation (RAG).

- Report or work order generation for network OAMP: The LLM excels at interpreting natural language requests to retrieve and structure data. For example, it translates a user's request like "Show me all high-severity alarms from the last 24 hours" into a formal SQL query to be executed against a database. Once the data is retrieved, the LLM then summarizes the results and format the information into a human-readable report or a structured work order. The final generation of formatted files (e.g. PDF) is handled by separate tools that use the LLM's text output as input.

- Solution analysis and generation for network OAMP: The LLM acts as an analytical assistant, helping engineers diagnose complex network problems. By processing information from its training and real-time data provided via RAG (such as logs, alarms, and performance metrics), the LLM generates hypotheses about the root cause of an issue or suggest potential troubleshooting strategies. These suggestions are intended to inform the decisions of human experts, not to be executed automatically.

- Intelligent task scheduling for network OAMP: The LLM functions as a planning engine by parsing high-level objectives described in natural language (e.g. "Investigate poor performance for customer X"). It then decomposes this objective into a logical sequence of sub-tasks required for the investigation. This generated plan is then passed to specialized automation systems or human operators for validation and execution within the OAMP infrastructure.

6.2 Knowledge-based Q&A

6.2.1 Knowledge Q&A

6.2.1.1 Description

Network OAMP staff rely on interconnected technical documentation for operations. A RAG architecture, along with appropriate document representation methods that preserve relationships and hierarchies, assists in retrieving relevant information. To improve the accuracy and robustness of this process, the LLM is prepared using advanced fine-tuning methods like Retrieval-Augmented Fine-Tuning (RAFT), which trains the model to better discern relevant facts from irrelevant information. The LLM then generates more reliable and accurate responses based on this retrieved information, helping staff locate and understand relevant technical information more efficiently.

6.2.1.2 Process

Pre-condition: Network OAMP knowledge sources (e.g. technical documents, operational guides) have been ingested and indexed into a retrieval-focused **Support System** (such as a vector database).

a) The user submits a query in natural language, such as "what are the causes of network congestion and how to solve it quickly" to the AI-Enhanced Network OAMP System.

b) The AI-Enhanced Network OAMP System, acting as the orchestrator, sends the user's query to the appropriate Support System to perform the retrieval step of the RAG process.

c) The Support System searches its indexed knowledge base and returns the most relevant documents or data snippets to the AI-Enhanced Network OAMP System.

d) The AI-Enhanced Network OAMP System constructs a new, detailed prompt. This prompt includes the original user query along with the retrieved documents, providing the necessary context for the LLM to generate a grounded answer.

e) The AI-Enhanced Network OAMP System sends this complete prompt in an API call to the Network OAMP LLM Service.

f) The Network OAMP LLM generates a response by reasoning over the provided query and context. The Network OAMP LLM Service returns this generated text to the AI-Enhanced Network OAMP System.

g) The AI-Enhanced Network OAMP System performs any final formatting on the response and delivers it to the user.

Table 6-1: Network OAMP Q&A

Details of the network OAMP LLM		
Basic Information	Model Classification	Language
	Model Usage	Network OAMP Q&A
Pre-training	Data Sources	Data in the telecommunication industry that has been appropriately cleansed and anonymised, professional technical documents and manuals for network OAMP operators.
	Data Modality	Text
Fine-tuning	Data Sources	Domain-specific Q&A pairs covering network operations scenarios, context for each Q&A pair to ensure accurate responses, and examples of different query types. Data types include network telemetry data, configuration files, log files, operational procedures, troubleshooting guides, and network documentation.
	Dataset Structure	Data needs to preserve relationships between documents, technical documentation hierarchies need to be maintained, and cross-references between related documents need to be preserved.
	Fine-tuning Methodologies	To improve performance within a RAG architecture, advanced methods like Retrieval-Augmented Fine-Tuning (RAFT) are needed. RAFT trains the model to reason over retrieved documents and explicitly ignore irrelevant "distractor" information, thereby reducing hallucinations and improving robustness.
	Data Modality	Text
Inference	Input	A detailed prompt constructed by the AI-Enhanced Network OAMP System. This prompt contains the original user query combined with the factual context retrieved from knowledge sources via the RAG process.
	Knowledge Enhancement	RAG is the core technique used at inference. The AI-Enhanced Network OAMP System retrieves relevant, up-to-date information from external knowledge bases and provides it to the LLM as context. This process grounds the model's response in facts, significantly improving accuracy and relevance. Prompt Engineering is integral to making the RAG process effective. A sophisticated prompt is engineered to instruct the LLM on precisely how to use the retrieved context, synthesize information from multiple sources, and formulate a coherent answer based on the user's original query.

Post-condition: The user receives a response that is synthesized by the LLM and grounded in the relevant information retrieved from the knowledge sources.

6.2.2 Core Network Signalling Anomaly Analysis

6.2.2.1 Description

Analysing signalling traffic in a 5G Core network to diagnose issues presents significant technical challenges, including a multitude of network interfaces, complex protocol interactions, and intricate service flows. This complexity demands deep expertise and makes manual analysis inefficient.

This application scenario uses a multi-stage, *hybrid AI architecture* to accelerate the diagnostic process. The architecture relies on a clear separation of concerns:

- Specialized Support Systems: These are dedicated anomaly detection engines and protocol analyzers responsible for the initial, heavy-lifting analysis. They ingest and process high-volume, raw signalling traces to identify statistically significant deviations from baseline behavior.
- The AI-Enhanced Network OAMP System: This system orchestrates the end-to-end workflow. When a support system detects an anomaly, the OAMP System retrieves the structured output (e.g. an alert detailing the affected network functions, interfaces, and specific protocol messages).
- The Network OAMP LLM: The role of the LLM is not to analyze the raw data, but to act as a high-level reasoning and synthesis engine. The OAMP System provides the LLM with the structured anomaly data from the support systems. The LLM then performs several key tasks:
  - Synthesizes Findings: It translates the cryptic, technical details of the anomaly into a clear, natural language summary.
  - Provides Context: Using Retrieval-Augmented Generation (RAG), it queries knowledge bases for related information, such as historical incident reports, technical documentation for the involved protocols, or known bug reports, providing valuable context to the operator.
  - Generates Documentation: It automatically drafts a structured summary of the event, including the synthesized findings and contextual information, which is used to populate a work order or incident ticket.

This approach leverages the LLM for its strengths in language and reasoning, while relying on specialized tools for the initial data processing. This significantly reduces the manual effort required to diagnose issues and process complaint work orders.

6.2.2.2 Process

Pre-condition: Specialized Support Systems (e.g. protocol analysers, anomaly detection engines) are in place to process raw signalling data. Historical incident reports, technical documentation, and root cause analyses have been ingested and indexed into a knowledge base for use by a Retrieval-Augmented Generation (RAG) system.

Users send the signalling detail list to the network OAMP LLM service:

a) The user initiates an analysis request via the AI-Enhanced Network OAMP System, providing an identifier for the relevant signalling trace or event data.

b) The AI-Enhanced Network OAMP System orchestrates the workflow. It first routes the raw signalling data to a specialized Support System (e.g. an anomaly detection engine) for initial processing and analysis.

c) The Support System analyzes the data and returns a structured output to the OAMP System. This output contains the technical details of the detected anomaly (e.g. affected network functions, specific protocol message deviations, timestamps).

d) The AI-Enhanced Network OAMP System then uses the RAG technique. It queries its knowledge base for historical cases, documentation, or known issues that match the characteristics of the detected anomaly.

e) The OAMP System constructs a detailed prompt for the LLM. This prompt includes the structured anomaly data from the support system and the relevant contextual documents retrieved via RAG.

f) The OAMP System sends this complete prompt in an API call to the Network OAMP LLM Service.

g) The Network OAMP LLM synthesizes the provided information and generates a response. This response includes a human-readable summary of the fault, a hypothesis about the root cause based on the combined data, and references to the evidence found in the retrieved documents.

h) The AI-Enhanced Network OAMP System receives the LLM's response, formats it into a final report along with the raw technical findings, and delivers it to the user.

Table 6-2: LLM assisted core network abnormal signalling

Details of LLM		
Basic Information	Model Classification	Language
	Model Usage	Synthesizing structured anomaly data, providing context from historical cases, and generating human-readable summaries and root cause hypotheses.
Pre-training	Data Sources	General telecommunication industry data, professional technical documents, and anonymised network operations data.
	Data Modality	Text
Fine-tuning	Data Sources	Fine-tuning data consists of structured anomaly reports paired with their corresponding human-written root cause analyses, historical incident tickets, and technical documentation explaining signalling protocols and network behavior.
	Data Modality	Text
Inference	Input	A detailed prompt constructed by the AI-Enhanced Network OAMP System. This prompt contains the structured output from a specialized anomaly detection engine, plus relevant historical cases and documentation retrieved via RAG.
	Knowledge Enhancement	Hybrid AI Approach: The system relies on specialized, non-LLM Support Systems for initial, high-volume signalling analysis and anomaly detection. The LLM's role is synthesis, not initial detection. Retrieval-Augmented Generation (RAG): After an anomaly is detected, RAG is used to retrieve relevant historical cases and technical documentation to provide context for the LLM's reasoning process.

Post-condition: The user receives a structured analysis report containing the technical findings from the specialized support systems, a natural language summary generated by the LLM that explains the anomaly and its likely cause, and links to supporting evidence from the knowledge base.

6.3 Report or Work Order Generation

6.3.1 Data query for network OAMP

6.3.1.1 Description

Network OAMP staff often need to refer to various data indicators as part of their tasks. This scenario simplifies the task of data querying by using an LLM as a front-end to enable the user to express what they want in natural language. The LLM is utilized to generate an appropriate data query (e.g. a set of SQL statements) and corresponding parameters. *Before the query is allowed to run, it is validated by a qualified network engineer*. After the query has run, the Network OAMP LLM generates a natural language summary of the data. The OAMP System receives this summary, performs any final formatting, and delivers the final answer to the user.

6.3.1.2 Process

Pre-condition: Relevant database schema descriptions or API documentation have been ingested and indexed into a vector database for retrieval. The system connects to the data source using a dedicated, read-only user account with the minimum required privileges:

a) The user describes the question in natural language, such as "What is the number of 5G users of China at 10:30 in October 2023" to the AI-Enhanced Network OAMP System.

b) The AI-Enhanced Network OAMP System uses RAG to retrieve the relevant database schema or API documentation from its knowledge base to understand how to query the requested data.

c) The OAMP System constructs a detailed prompt containing the user's query and the retrieved schema/API documentation. It sends this prompt in an API call to the Network OAMP LLM Service.

d) The Network OAMP LLM generates a structured query (e.g. a set of SQL statements) and returns it to the OAMP System.

e) The AI-Enhanced Network OAMP System passes the LLM-generated query to a Security Validation Engine. This engine parses the query and rejects it if it contains any forbidden or destructive commands (e.g. UPDATE, DELETE, DROP). Only safe, read-only queries are allowed to proceed.

NOTE: Enabling *any* set of SQL statements to be issued is beyond the scope of the present document.

f) Upon successful validation, the OAMP System executes the query against the appropriate database or API.

g) The OAMP System receives the query results (e.g. a data table). It then constructs a second prompt containing these results and an instruction to summarize them. It sends this prompt to the Network OAMP LLM Service.

h) The Network OAMP LLM generates a natural language summary of the data. The OAMP System receives this summary, performs any final formatting, and delivers the answer to the user.

Table 6-3: Data query for network OAMP

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1) Translating natural language questions into structured queries (e.g. SQL, API calls). 2) Summarizing the structured data returned from the query into a natural language response.
Pre-training	Data Sources	General data, professional technical documents, and a wide corpus of code and structured query languages.
	Data Modality	Text
Fine-tuning	Data Sources	Fine-tuning data consists of natural language questions paired with their corresponding correct and safe queries (e.g. text-to-SQL or text-to-API-call examples).
	Data Modality	Text
Inference	Input	A detailed prompt containing the user's question and relevant context, such as database schema definitions or API documentation retrieved via RAG.
	Knowledge Enhancement (RAG)	RAG is used to retrieve the specific database table schemas or API documentation relevant to the user's query. This context is essential for the LLM to generate a syntactically correct and executable query.
Security	Query validation	**This is a system-level requirement, not an LLM feature**. Every query generated by the LLM is processed by a security sandbox that validates it against a strict allow-list of read-only commands before execution to prevent data modification, deletion, or injection attacks.

Post-condition: The user receives a natural language answer to their question, supported by the data retrieved from the database or API.

6.4 Solution analysis and generation

6.4.1 Network performance assurance

6.4.1.1 Description

Ensuring network performance for major, high-visibility events requires the complex coordination of monitoring, analysis, and proactive optimization tasks. The goal of using an LLM in this scenario is to assist human experts by acting as a powerful planning and synthesis engine. The LLM helps accelerate the creation of a comprehensive assurance plan by interpreting the high-level requirements of an event, decomposing it into necessary operational tasks, and synthesizing data from various specialized monitoring and analysis systems. The final plan is a recommendation that is *validated and approved by a qualified engineer*.

6.4.1.2 Process

Pre-condition: The system has access to relevant API descriptions for specialized support systems. Historical assurance plans, network documentation, and performance baselines have been indexed in a knowledge base for RAG:

a) A user submits a high-level request for a network assurance plan (e.g. for the Hangzhou Asian Games opening ceremony) to the AI-Enhanced Network OAMP System.

b) The AI-Enhanced Network OAMP System, acting as the orchestrator, uses RAG to retrieve relevant context from its knowledge base, such as past assurance plans for similar large-scale events.

c) The OAMP System sends a detailed prompt to the Network OAMP LLM Service. The prompt includes the user's request and the retrieved contextual documents.

d) The Network OAMP LLM generates a high-level task flow (a proposed plan) for creating the assurance solution. This plan identifies the necessary steps, such as identifying critical network elements, establishing performance monitoring dashboards, and defining potential optimization actions.

e) The AI-Enhanced Network OAMP System then executes the analytical steps of the plan by making API calls to specialized Support Systems. For example:

   a. it calls a Performance Monitoring System to gather baseline data and identify key performance indicators (KPIs) for the specified location and services;

   b. it calls a Topology and Asset Management System to identify all critical network infrastructure supporting the event;

   c. it optionally calls a Root Cause Analysis Engine to identify potential points of failure.

f) The OAMP System gathers the structured data from these support systems and constructs a final, detailed prompt for the LLM. This prompt asks the LLM to synthesize all the information into a comprehensive assurance plan document.

g) The Network OAMP LLM generates the draft assurance plan, including recommended configurations, monitoring thresholds, and contingency actions.

h) The AI-Enhanced Network OAMP System presents the complete draft plan to a qualified human engineer for review. The engineer validates the plan. If there is a problem, the engineer modifies it and then *give explicit approval before any part of the plan is implemented in the live network.*

NOTE: While it is possible for the network engineer to resubmit the plan to the Network OAMP System to help teach it, this is beyond the scope of the present document.

Table 6-4: Network performance assurance

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1) Task Decomposition: Breaking down a high-level assurance request into a logical sequence of planning and analysis steps. 2) Data Synthesis: Consolidating and summarizing structured data from multiple specialized support systems into a coherent, human-readable assurance plan.
Pre-training	Data Sources	General data, professional technical documents, and anonymised network OAMP documentation.
	Data Modality	Text
Fine-tuning	Data Sources and Scale	Fine-tuning data includes historical network assurance plans, incident post-mortems, network configuration guides, and examples of decomposing high-level goals into specific operational tasks.
	Data Modality	Text
Inference	Input	A detailed prompt from the AI-Enhanced Network OAMP System containing the user's goal, plus structured data outputs from various support systems (e.g. performance metrics, asset lists).
	Knowledge Enhancement (RAG)	**Hybrid AI Architecture**: The LLM does not perform analysis directly. It acts as a planning and synthesis layer that orchestrates calls to and interprets results from specialized, non-LLM Support Systems (e.g. monitoring platforms, analysis engines). RAG is used to retrieve historical assurance plans and best-practice documents to provide the LLM with relevant examples and context for generating the new plan.
Safety	Human-in-the-Loop Oversight	**This is a mandatory system-level requirement.** *Any plan or configuration change generated by the LLM that could affect the live network is presented to a qualified human engineer for validation, modification, and explicit approval before execution.*

Post-condition: A qualified network engineer receives a comprehensive, LLM-generated draft assurance plan for review, modification and final approval.

6.5 Intelligent task scheduling

6.5.1 Network fault monitoring

6.5.1.1 Description

In network fault monitoring, an LLM serves as a powerful synthesis and reasoning engine to assist human operators, not replace them. Within a hybrid AI architecture for network fault monitoring, the LLM's primary role is to translate and summarize the structured, technical outputs from specialized fault detection and analysis systems into human-readable language. By processing alarms and logs and correlating them with historical data using RAG, the LLM generates a concise fault summary, propose a root cause hypothesis, and recommend a sequence of diagnostic or remedial actions. These recommendations are presented to a qualified engineer for validation and approval before any action is taken.

6.5.1.2 Process

Pre-condition: Specialized Support Systems for fault detection and root cause analysis are operational. Historical fault cases and technical documentation are indexed in a knowledge base for RAG:

a) An alarm is ingested by the AI-Enhanced Network OAMP System, or a user initiates a fault investigation.

b) The OAMP System, acting as the orchestrator, routes the initial alarm and associated data to a specialized Support System (e.g. a Fault Detection Engine) for primary analysis.

c) The Support System returns structured data identifying the fault and its immediate parameters to the OAMP System.

d) The OAMP System optionally calls a second Support System (e.g. a Root Cause Analysis Engine) with this enriched data to perform deeper analysis and pinpoint potential causes.

e) The OAMP System uses RAG to retrieve relevant historical incident reports and technical documents from its knowledge base that match the current fault signature.

f) The OAMP System constructs a detailed prompt containing the structured outputs from the support systems and the contextual documents from RAG. It sends this prompt in an API call to the Network OAMP LLM Service.

g) The Network OAMP LLM synthesizes all the provided information to generate a response that includes:

   1) A clear, natural language summary of the fault.

   2) A hypothesis on the most likely root cause.

   3) A recommended, step-by-step action plan for resolution.

h) The AI-Enhanced Network OAMP System presents the full report - including the LLM's summary and recommended plan - to a qualified human operator. The operator reviews, validates, and explicitly approves the action plan before proceeding.

i) The operator executes the approved remedial actions using the OAMP system's tools.

Table 6-5: Network fault monitoring

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1. Data Synthesis: Summarizing and translating structured technical data from multiple support systems into a human-readable fault description. 2. Hypothesis Generation: Proposing a likely root cause based on current and historical data. 3. Task Planning: Generating a recommended sequence of diagnostic or remedial actions for human review.
Pre-training	Data Sources	General data, professional technical documents, and anonymised network OAMP documentation.
	Data Modality	Text
Fine-tuning	Data Sources	Fine-tuning data includes historical incident reports, alarm-to-root-cause mappings, and successful resolution procedures from past network faults.
	Data Modality	Text
Inference	Input	A detailed prompt from the AI-Enhanced Network OAMP System containing structured data from fault detection/analysis engines and relevant historical cases retrieved via RAG.
	Knowledge Enhancement (RAG)	**Hybrid AI Architecture**: The LLM does not perform initial fault detection. It synthesizes the outputs from specialized, non-LLM Support Systems that are designed for high-speed, accurate anomaly detection and analysis. RAG is used to retrieve similar historical fault cases and technical documentation to provide the LLM with context for its root cause hypothesis and action plan recommendation.
Safety	Human-in=the-Loop Oversight	**This is a mandatory system-level requirement.** All diagnoses, root cause hypotheses, and especially any recommended configuration changes or remedial actions generated by the LLM are first presented to a qualified human operator for validation and explicit approval before execution.

Post-condition: A human operator receives a comprehensive fault report containing a summary, a root cause hypothesis, and a recommended, validated action plan for final approval and execution.

6.5.2 RAN optimization

6.5.2.1 Description

Radio Access Network (RAN) optimization is a highly complex task requiring deep domain expertise. An LLM serves as a powerful assistant to a RAN engineer by automating the initial stages of analysis and solution drafting. The LLM's role is to interpret optimization work orders, decompose the problem into a logical sequence of analytical tasks, and synthesize the outputs from specialized RAN monitoring and analysis systems. It then generates a draft optimization report, including a root cause hypothesis and a recommended solution (e.g. a specific parameter change). *This draft is presented to a qualified engineer for rigorous validation, potential modification, and final approval before any action is taken on the live network due to the chance of adversely affective network users.*

6.5.2.2 Process

Pre-condition: The system is integrated with specialized RAN Support Systems. Historical optimization cases and RAN technical documentation are indexed in a knowledge base for RAG:

a) A user submits a RAN optimization task (e.g. a work order ID) to the AI-Enhanced Network OAMP System.

b) The OAMP System, serving as the orchestrator, uses RAG to retrieve the work order details and any relevant historical optimization cases from its knowledge base.

c) The OAMP System calls a specialized Support System (e.g. a RAN Performance Monitoring Platform) to gather real-time and historical performance data (KPIs) for the specified cell(s).

d) The OAMP System optionally sends this performance data to another Support System (e.g. a Root Cause Analysis Engine) to identify the most likely technical causes for the performance degradation.

e) The OAMP System constructs a detailed prompt for the LLM. This prompt includes the work order intent, the structured data from the support systems, and the context from the RAG retrieval.

f) The OAMP System sends the prompt in an API call to the Network OAMP LLM Service.

g) The Network OAMP LLM synthesizes all the provided information and generates a draft optimization report containing:

   1) A summary of the problem.

   2) A hypothesis on the root cause.

   3) A recommended, specific optimization action (e.g. "Recommend adjusting parameter X to value Y").

h) The AI-Enhanced Network OAMP System presents the complete draft report and the specific recommended action to a qualified human RAN engineer. The engineer uses his or her expert judgment to validate the analysis, confirm the safety and efficacy of the proposed change, and provide explicit approval before the change is implemented.

Table 6-6: Large language model assisted RAN optimization

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1. Task Decomposition: Interpreting a work order and planning the analysis sequence. 2. Data Synthesis: Summarizing structured KPI data and analysis results from support systems. 3. Solution Recommendation: Generating a draft optimization solution for expert human review.
Pre-training	Data Sources	Operators' internal specifications, RAN optimization cases, product manuals, and other relevant technical corpora.
	Data Modality	Text
Fine-tuning	Data Sources	Fine-tuning data includes RAN knowledge Q&A corpora, technical texts, historical optimization work orders paired with their successful solutions, and cell configuration data.
	Data Modality	Text
Inference	Input	A detailed prompt from the AI-Enhanced Network OAMP System containing the work order, structured data from RAN analysis tools, and relevant historical cases retrieved via RAG.
	Knowledge Enhancement (RAG)	**Hybrid AI Architecture**: The LLM does not analyse raw RAN data. It orchestrates and synthesizes outputs from specialized Support Systems (e.g. performance monitoring platforms, RCA engines) that perform the primary analysis. RAG is used to retrieve similar historical optimization cases and technical documentation to provide the LLM with context for generating its recommendation.
Safety	Human-in-the-Loop Oversight	**This is a mandatory system-level requirement.** *Any optimization solution or configuration change recommended by the LLM is presented to a qualified human RAN engineer for validation and explicit approval before being implemented in the live network.*

Post-condition: A qualified RAN engineer receives a comprehensive draft optimization report, including a root cause hypothesis and a recommended action plan, for their expert review and final decision-making.

6.5.3 Core network alarm diagnosis

6.5.3.1 Description

Diagnosing alarms in a 5G core network is a complex task requiring the correlation of data from numerous network functions and infrastructure layers. An LLM serves as a powerful assistant to an expert operator by automating the process of data gathering and synthesis. The LLM's role is to interpret an alarm, orchestrate calls to various specialized monitoring and analysis systems, and then consolidate the findings into a single, human-readable report. The present document includes a summary of the issue, a root cause hypothesis, and a recommended diagnostic or remedial action plan. *This entire output is treated as a draft for expert human review.*

6.5.3.2 Process

Pre-condition: The system is integrated with specialized Support Systems for core network monitoring, topology mapping, and root cause analysis. Historical alarm tickets and technical documentation are indexed in a knowledge base for RAG:

a) An alarm is ingested, or a user initiates a diagnostic task (e.g. "diagnose alarm: destination NF service unreachable on NE X") in the AI-Enhanced Network OAMP System.

b) The OAMP System, as the orchestrator, calls multiple specialized Support Systems to gather relevant data. This includes:

   1) A Monitoring System to get performance metrics and health status from the affected NE, its underlying VMs, and hosts.

   2) A Topology System to understand the service chain and dependencies related to the alarm.

3) A Log Analysis System to pull relevant logs from the involved components.

c) The OAMP System also uses RAG to retrieve historical incident reports and documentation related to this specific alarm type and NE model.

d) The OAMP System constructs a detailed prompt containing the structured data from all support systems and the contextual documents from RAG. It sends this prompt in an API call to the Network OAMP LLM Service.

e) The Network OAMP LLM synthesizes all the provided information to generate a draft diagnostic report containing:

1) A clear summary of the fault.

2) A root cause hypothesis based on the correlated data.

3) A recommended, step-by-step action plan for resolution.

f) The AI-Enhanced Network OAMP System presents the complete draft report to a qualified human operator. The operator uses his or her expert judgment to validate the findings and explicitly approve the action plan before any remedial action is taken on the core network.

Table 6-7: Large language model assisted core network alarm diagnosis

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1. Data Synthesis: Consolidating and summarizing structured data from multiple core network support systems. 2. Hypothesis Generation: Proposing a likely root cause for an alarm based on correlated data. 3. Solution Recommendation: Generating a draft diagnostic and remedial plan for expert human review.
Pre-training	Data Sources	ICT domain general corpus, enterprise internal documents, network protocol specifications.
	Data Modality	Text
Fine-tuning	Data Sources	Fine-tuning data includes historical alarm tickets, successful resolution procedures, root cause analysis reports, and technical Q&A pairs for core network issues.
	Data Modality	Text
Inference	Input	A detailed prompt from the AI-Enhanced Network OAMP System containing structured data from monitoring/analysis engines and relevant historical cases retrieved via RAG.
	Knowledge Enhancement (RAG)	**Hybrid AI Architecture**: The LLM does not perform primary alarm analysis. It synthesizes outputs from specialized Support Systems (e.g. monitoring platforms, RCA engines) that handle the initial data processing. RAG is used to retrieve similar historical alarm tickets and technical documentation to provide context for the LLM's diagnosis.
Safety	Human-in-the-Loop Oversight	**This is a mandatory system-level requirement.** *Any diagnostic conclusion or recommended solution generated by the LLM is presented to a qualified human operator for validation and explicit approval before any action is taken on the live core network.*

Post-condition: A qualified network operator receives a comprehensive diagnostic report, including a root cause hypothesis and a recommended action plan, for their expert validation and final approval.

6.5.4 Slicing Packet Network troubleshooting

6.5.4.1 Description

Troubleshooting a Slicing Packet Network (SPN) is a critical and complex task. An LLM serves as a powerful assistant to a transport network engineer by automating the laborious process of data collection, correlation, and summarization. The LLM's role is to interpret a fault report, orchestrate API calls to specialized SPN monitoring and diagnostic tools, and synthesize the results into a coherent troubleshooting report. The present document presents a root cause hypothesis and a recommended action plan. This output is treated as a draft that requires rigorous validation by a qualified human expert before any action is taken.

6.5.4.2 Process

Pre-condition: The system is integrated with specialized Support Systems for SPN monitoring, path tracing, and performance analysis. Historical troubleshooting guides and incident reports are indexed in a knowledge base for RAG:

a) A user initiates a troubleshooting task (e.g. "diagnose the reason for base station 123.123.1.2 going offline") in the AI-Enhanced Network OAMP System.

b) The OAMP System, as the orchestrator, calls specialized Support Systems to gather data. This includes:

   1) An SPN Controller/NMS to check the status of the specified endpoint and its associated paths.

   2) A Performance Monitoring System to retrieve relevant KPIs along the service path.

   3) A Path Trace Tool to identify the exact route and any potential breaks.

c) The OAMP System uses RAG to retrieve historical incident reports and technical documentation related to similar SPN faults.

d) The OAMP System constructs a detailed prompt containing the structured data from all support systems and the contextual documents from RAG. It sends this prompt in an API call to the Network OAMP LLM Service.

e) The Network OAMP LLM synthesizes all the provided information to generate a draft troubleshooting report containing:

   1) A clear summary of the fault.

   2) A root cause hypothesis based on the correlated data.

   3) A recommended, step-by-step action plan for resolution.

f) The AI-Enhanced Network OAMP System presents the complete draft report to a qualified human engineer. The engineer uses his or her expert judgment to validate the findings and explicitly approve the action plan before any diagnostic or remedial action is taken on the live SPN.

Table 6-8: Large language model assisted SPN troubleshooting

Details of large language model		
Basic Information	Model Classification	Language
	Model Usage	1. Data Synthesis: Consolidating and summarizing structured data from specialized SPN support systems. 2. Hypothesis Generation: Proposing a likely root cause for a fault based on correlated data. 3. Solution Recommendation: Generating a draft troubleshooting and repair plan for expert human review.
Pre-training	Data Sources and Scale	ICT domain general corpus, enterprise internal documents, transport network technical specifications
	Data Modality	Text
Fine-tuning	Data Sources and Scale	Fine-tuning data includes historical SPN trouble tickets, successful resolution procedures, root cause analysis reports, and technical Q&A pairs for transport network issues.
	Data Modality	Text
Inference	Input	A detailed prompt from the AI-Enhanced Network OAMP System containing structured data from SPN monitoring/analysis tools and relevant historical cases retrieved via RAG.
	Knowledge Enhancement (RAG)	**Hybrid AI Architecture**: The LLM does not perform primary fault analysis. It synthesizes outputs from specialized Support Systems (e.g. NMS/controllers, path trace tools) that handle the initial data collection and analysis. RAG is used to retrieve similar historical trouble tickets and technical documentation to provide context for the LLM's diagnosis.
Safety	Human-in-the-Loop Oversight	**This is a mandatory system-level requirement.** *Any diagnostic conclusion or recommended solution generated by the LLM is presented to a qualified human engineer for validation and explicit approval before any action is taken on the live transport network.*

Post-condition: A qualified transport network engineer receives a comprehensive troubleshooting report, including a root cause hypothesis and a recommended action plan, for their expert validation and final approval.

7 Interaction processes in large language model assisted network OAMP

The interaction process for an LLM-assisted network OAMP system is an orchestrated workflow managed by the AI-Enhanced Network OAMP System. This system acts as a central controller, coordinating between the user, specialized support systems, and the Network OAMP LLM to achieve a user's goal safely and effectively:

1) User Request: The User submits a request in natural language to the AI-Enhanced Network OAMP System.

2) Orchestration and Tool Use (Iterative Loop): The AI-Enhanced Network OAMP System begins an iterative process to gather information and perform analysis. This typically involves multiple steps:

   a) It calls the appropriate Support System(s) (e.g. a monitoring platform, a database, a root cause analysis engine) via their APIs to retrieve structured data or perform specialized analysis.

   b) The Support System(s) return structured data to the OAMP System.

   c) The OAMP System repeats steps 2a-2b as many times as needed, calling different tools as needed in order to build a complete picture of the situation.

3) LLM Synthesis and Reasoning: Once sufficient data is gathered, the AI-Enhanced Network OAMP System constructs a detailed prompt. This prompt includes the original user request, all the structured data gathered from the Support Systems, and any relevant context retrieved via RAG. It sends this prompt in an API call to the Network OAMP LLM Service.

4) LLM Response: The Network OAMP LLM synthesizes the information and returns a generated response to the OAMP System. This response contains a summary, a root cause hypothesis, and/or a recommended action plan.

5) Validation and Safety Checks (Mandatory): The AI-Enhanced Network OAMP System processes the LLM's output through mandatory safety gates:

   a) Security Validation: If the output is a query or command (e.g. SQL), it is checked by a validation engine to ensure it is safe and read-only.

   b) Human-In-The-Loop (HITL) Approval: If the output is a recommendation that could alter the state of the live network, it is first presented to a qualified human operator for expert review, modification, and explicit approval before it is declared ready to be deployed.

6) Only after all necessary validation and approval steps are complete does the AI-Enhanced Network OAMP System deliver the final, verified response or execute the approved action.

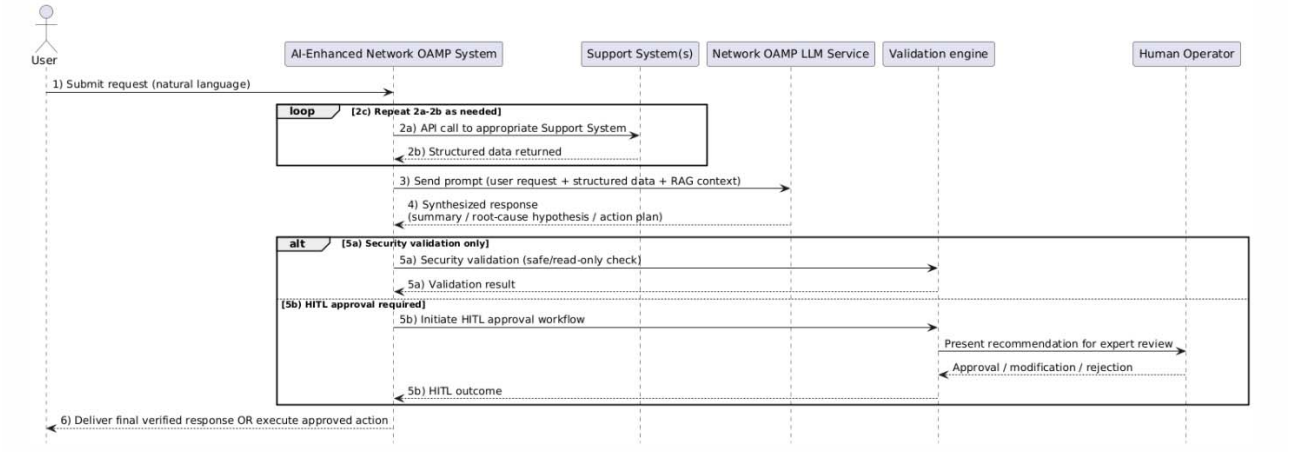

Figure 7-1: Interaction processes

8 Standardization Recommendation for an AI-Enhanced Network OAMP System

The present document has provided a detailed study of typical application scenarios for an AI-Enhanced Network OAMP System. It has analysed several key use cases, including knowledge-based Q&A, report generation, solution analysis, and intelligent task scheduling. Through this analysis, a robust and safe architectural framework has been established, clarifying the distinct roles of the user-facing AI-Enhanced Network OAMP System, the backend Network OAMP LLM Service, and the specialized Support Systems.

The key findings of this study reveal that the effective and safe application of LLMs in network OAMP is not achieved by a single monolithic model, but through a Hybrid AI architecture. This architecture leverages the LLM primarily as a powerful engine for task decomposition, data synthesis, and reasoning, while relying on specialized, non-LLM support systems for primary data analysis and fault detection. This study concludes that for any scenario with the potential to alter the state of a live network, Human-In-The-Loop (HITL) validation is not an optional feature but a core, non-negotiable safety requirement.

Consequently, to move forward responsibly, a phased approach to standardization is recommended:

1) Establish Architectural and Safety Baselines: The immediate priority for standardization is to formally define the core architectural components and mandate fundamental safety principles. This includes standardizing the requirement for a clear separation of concerns between the orchestrating application and the LLM service and codifying the necessity of security validation sandboxes and HITL approval workflows for all safety-critical operations.

2) Develop Evaluation and Risk-Assessment Frameworks: It is recommended that future work focus on creating standardized frameworks for evaluating the performance and risk of these systems. This includes defining metrics to measure the accuracy of LLM-generated hypotheses and the prevalence of hallucinations, ensuring that all systems are rigorously and consistently benchmarked for safety and reliability before deployment.

3) Standardize Interfaces and Data Models: Standardization efforts for defining common APIs and data models only proceed after a safe architectural baseline and robust evaluation methods are established. This will ensure that interfaces between the AI-Enhanced Network OAMP System, the LLM Service, and various Support Systems are built upon a foundation of safety and interoperability.

History

Document history		
V4.1.1	October 2025	Publication