AmiEs 2026

AIM5B: AI Integrated Semantic Framework for 5G and Beyond Network Management

Thanveer Sulthana, Venkat Rao Manavarthi, Jayadithya Nalajala, Ankitha Srirama Reddy,
Baek-Young Choi, Sejun Song

Missouri University of Science and Technology, MO, USA

OUTLINE

- Motivation
- Problem Statement
- Proposed Solution
- System Architecture
- 5G Testbed Infrastructure
  - 5G Testbed & Data Collection
- Knowledge Graph Modelling
  - Semantic Transformation & Knowledge Graph Data Efficiency

Semantic Communication



Source Data

201,578 bytes

Explain the image.

Encoder (ChatGPT 4o DALL.E)

"A joyful young boy with short curly hair, wearing a light gray t-shirt and blue jeans. He is singing into a microphone with great enthusiasm, his face lit up with happiness. One hand holds the microphone while the other is clenched in a fist, showing his passion and energy. The stage setting in the background features a spotlight that highlights him, emphasizing his engaging performance and focus on singing.."

Semantic Data

429 bytes

create an image of ***(Semantic Data)***

Decoder (ChatGPT 4o DALL.E)

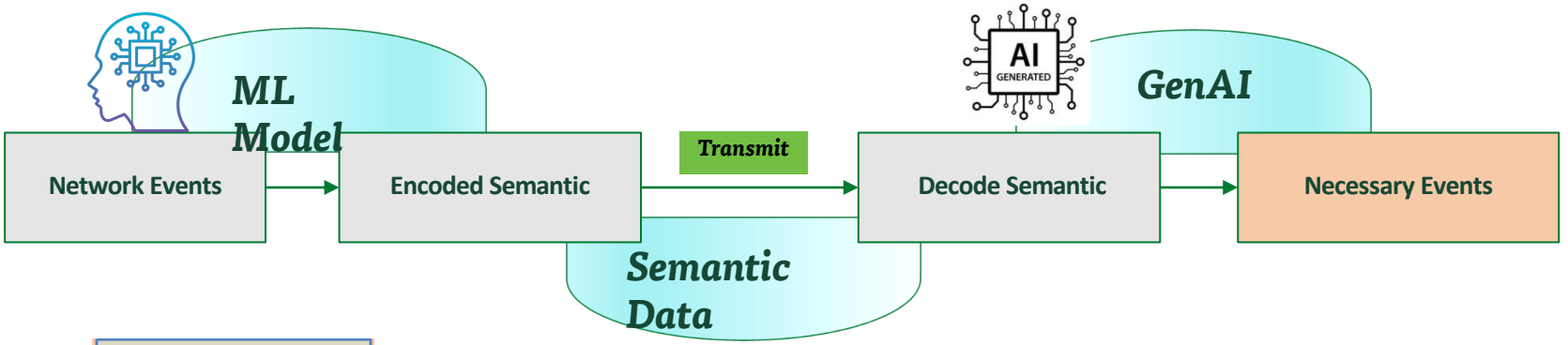


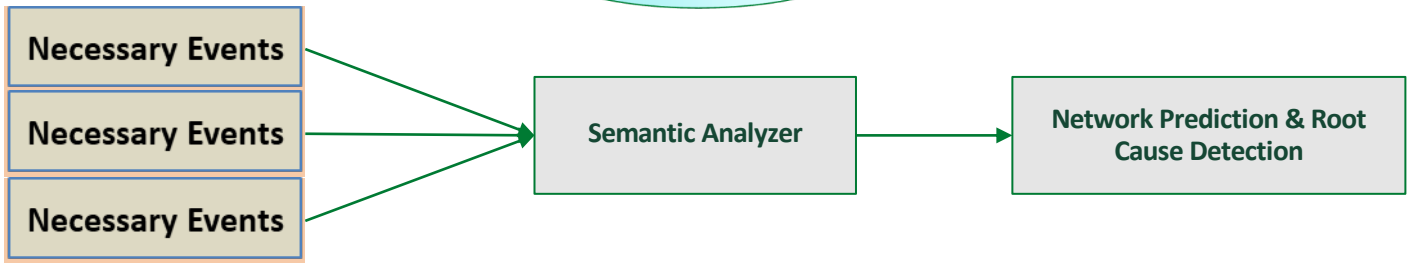
Target Data

SiNeMa: Semantic-Infused Network Management (in 2024 SFFP)

For AI-centric network, we need a new network management framework. Apply semantic communication to network management for reducing and securing messages.

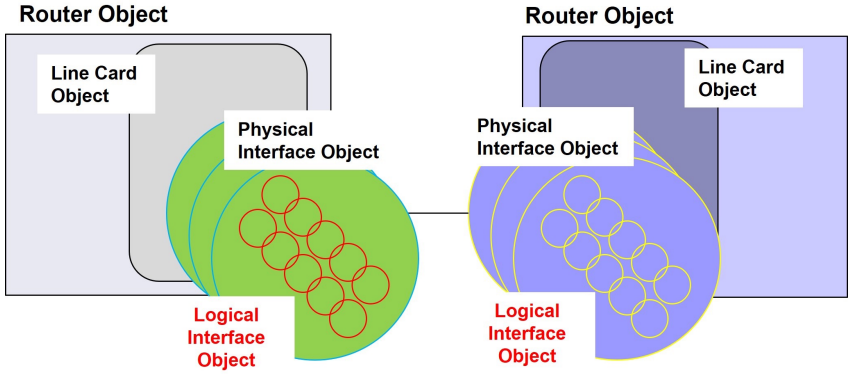
- Don't transmit but generate!
- Analyze in time.





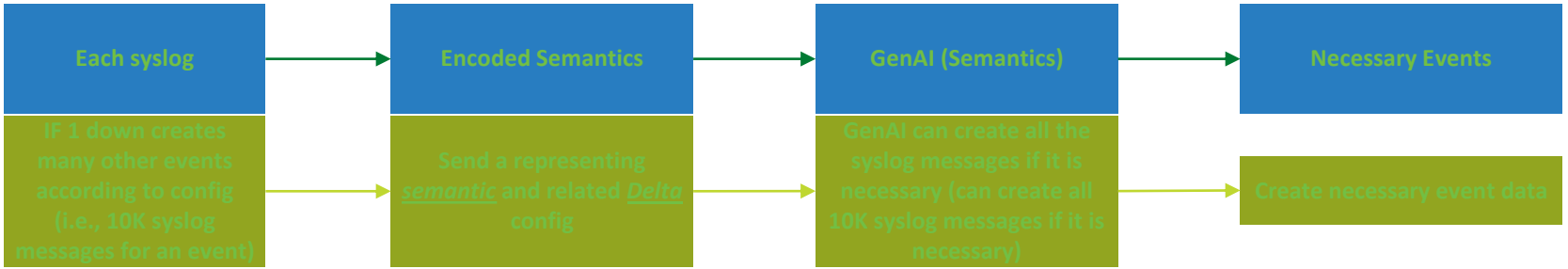
SiNeMa Case Study with Syslog

Campus network syslog



- Learning from 10 years of syslog and SNMP messages
- Encoding semantics
- Create a GenAI model for decoding
- Decoding from semantics
- Ground truth to validate accuracy

A single failure may cause multiple events and syslog due to the dependency (e.g., a physical line card can configure 100 virtual interfaces.)



OUTLINE

- AI-Driven Querying & Network Insights
  - Query Crafting (Manually and Code Iteration) & Grouping
- Enhancement Stages for Improving Accuracy
- Evaluation Metrics for Retrieval Performance
- Accuracy Performance Before and After Enhancement Stages
- Log Data Reduction and Efficiency
- System Performance and Throughput Analysis
- Conclusion & Future Work

Motivation

- 5G core functions produce large volumes of heterogeneous and interdependent control-plane logs.
- Traditional dashboards and static rule-based tools do not capture semantic context well.
- Network operators struggle to perform accurate and timely root cause analysis in complex operational environments. For Example, access to events such as registration, NF discovery, and PDU sessions.
- The challenge is to retrieve the right evidence with high precision from dense control-plane logs.
- AIM5B addresses this with semantic structuring and LLM-driven graph querying.

Problem Statement

- Existing solutions lack the ability to interpret control-plane semantics and correlate events across multiple network functions.
- Failures in 5G networks are inherently cross-layer and involve temporal dependencies between components such as AMF, NRF, and UPF.
- Current systems treat logs as isolated records, leading to inefficient diagnosis, increased manual effort, and high processing overhead.
- There is a need for a unified framework that enables structured reasoning over semantically rich network data.

Proposed Solution

AIM5B introduces an AI-integrated semantic framework for 5G network management.

The core idea is to transform unstructured syslog data into a structured knowledge graph aligned with control-plane event semantics.

The framework consists of:

- Semantic transformation layer to normalize and structure logs
- Knowledge graph (Neo4j) to represent events and relationships
- AI agent (LLM) to interpret natural language queries
- Cypher-based query engine for precise and deterministic retrieval

Continue...

Unlike traditional approaches, AIM5B enables:

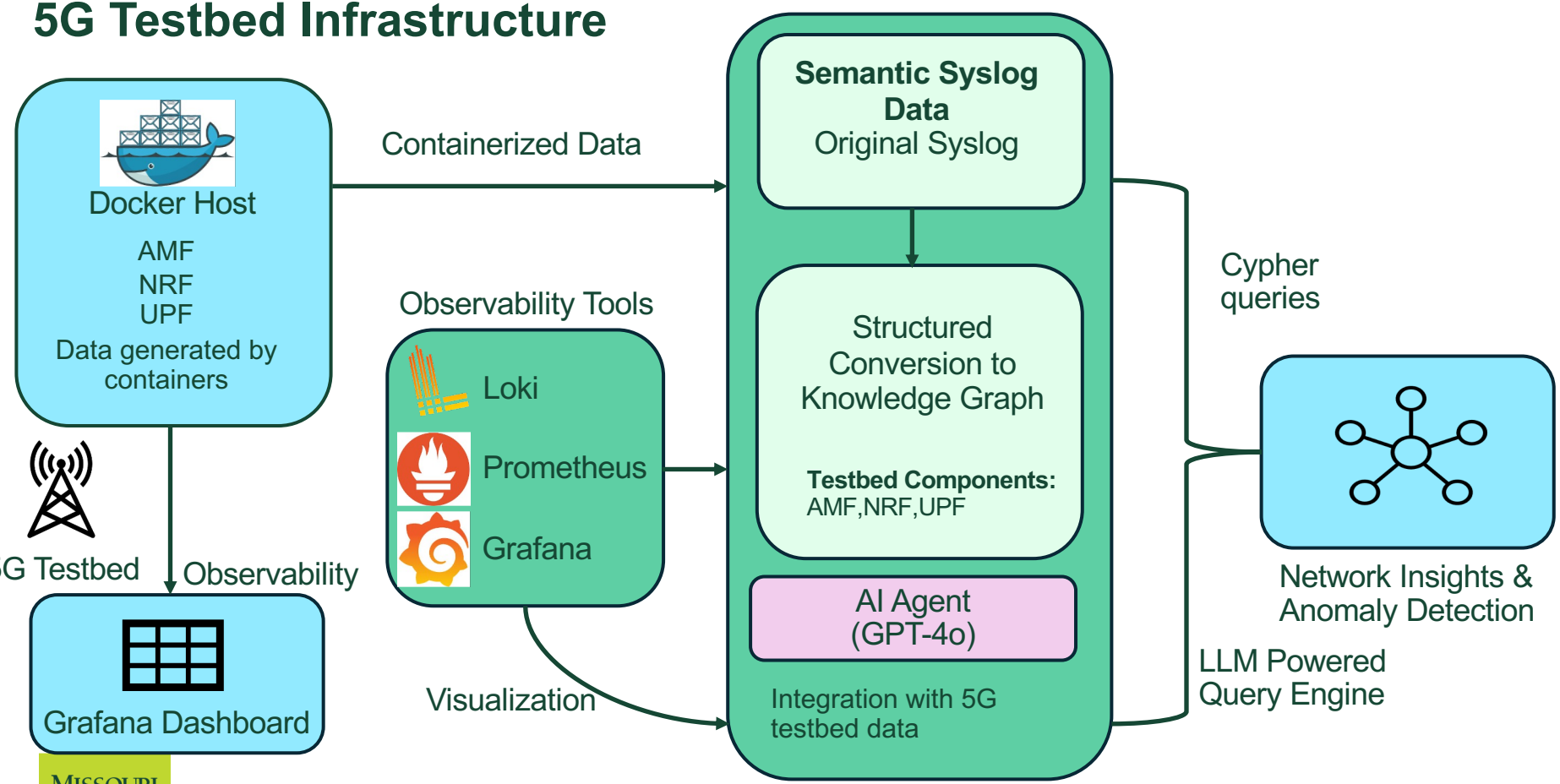
- Semantic, query-driven observability
- Event-level reasoning across components
- Explainable and structured analysis

The system is validated on a real private 5G testbed, ensuring practical relevance and scalability.

System Architecture

- The Aim5B architecture consists of a modular pipeline that ingests syslogs from a real 5G testbed
- Applies semantic transformation, constructs a knowledge graph using Neo4j, and enables query-driven analysis through an LLM-based agent.
- The system supports both real-time querying and scalable analysis of large log datasets.

5G Testbed Infrastructure



5G Testbed & Data Collection

- A private 5G testbed is deployed using containerized infrastructure (Docker) running core network functions: AMF, NRF, and UPF.
  - AMF(Access and Mobility Management Function)
  - NRF (Network Repository Function)
  - UPF (User Plane Function)
- These components continuously generate control-plane data representing key operations such as UE registration, NF discovery, and session management.
- The generated data is high-volume, unstructured, and distributed across multiple network functions, reflecting realistic network behavior.
- Observability tools are integrated for data collection:
  - Loki for log aggregation
  - Prometheus for metrics collection
  - Grafana for visualization
- While these tools enable monitoring and visualization, they do not provide semantic understanding or cross-component reasoning.

Core Network Components

AMF

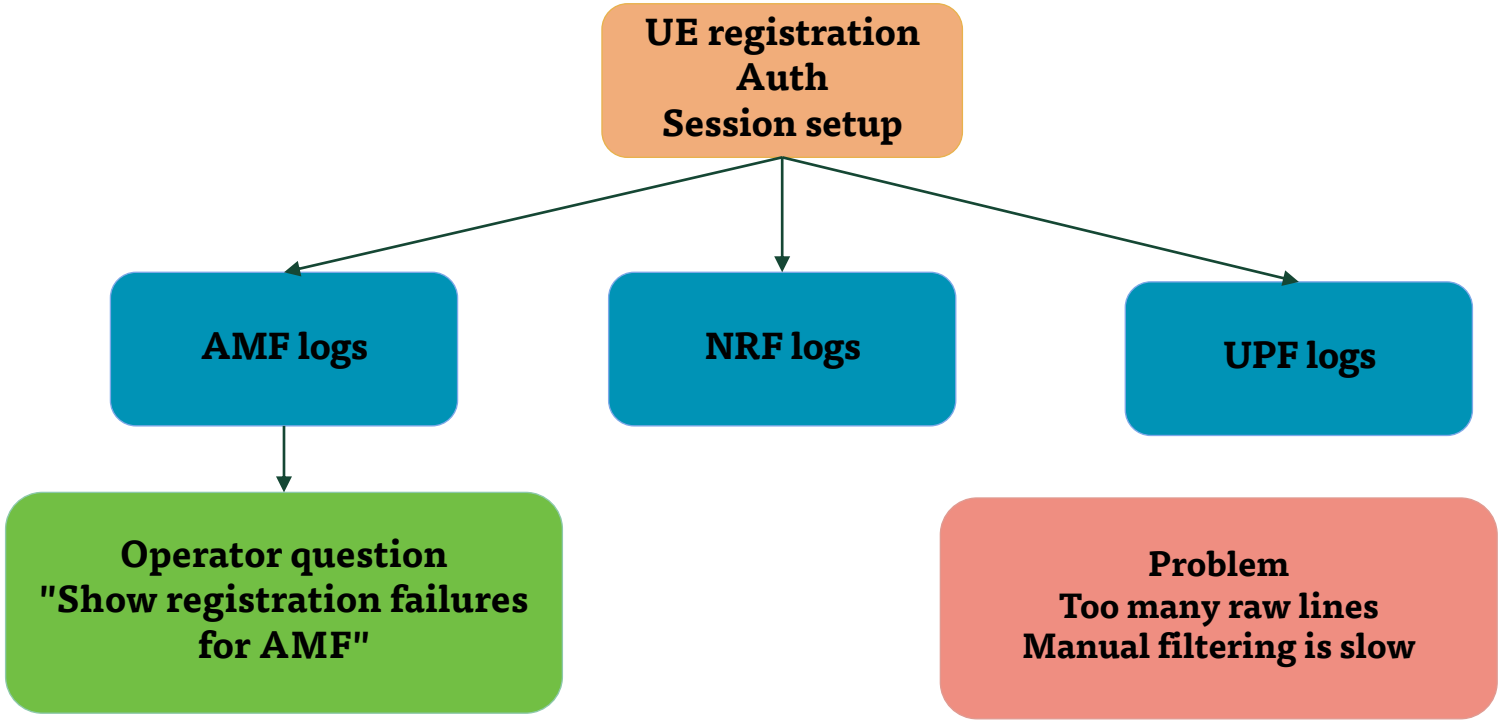
- Handles UE registration, authentication, mobility management
- Generates logs related to 5GMM state transitions, IMSI handling, and signaling (NGAP/SBI)

NRF

- Maintains network function registry and service discovery
- Logs include NF registration, discovery requests, heartbeat messages, and service status updates

UPF

- Manages data-plane routing, QoS(Quality of Service) enforcement, and session forwarding
- Logs capture PDU(Protocol Data Unit) session establishment, interface status, and data path events



AIM5B is motivated by a simple gap: operators can ask useful questions, but current tools do not translate those questions into precise telecom-aware retrieval over raw 5G logs.

Control-Plane Log Characteristics

Logs are:

- High-volume and unstructured
- Event-driven and time-sequenced
- Distributed across multiple network functions

Represent key procedures such as:

- UE registration and authentication
- NF discovery and service registration
- PDU session setup and release
- SBI (Service-Based Interface) interactions

Semantic Identifiers in Logs (structured)

- Each log entry contains explicit identifiers that define its semantic meaning:
- IMSI (International Mobile Subscriber Identity) → Identifies the user
- NF Instance ID → Identifies network functions
- Service Names → Identifies service modules (Network Function Management Service)
- SBI Transaction IDs → Tracks inter-component communication

Role in AIM5B

These identifiers act as semantic anchors for:

- Event correlation
- Query grounding
- Knowledge graph construction
- They enable AIM5B to perform precise, event-level reasoning instead of simple text matching

Key Insight:

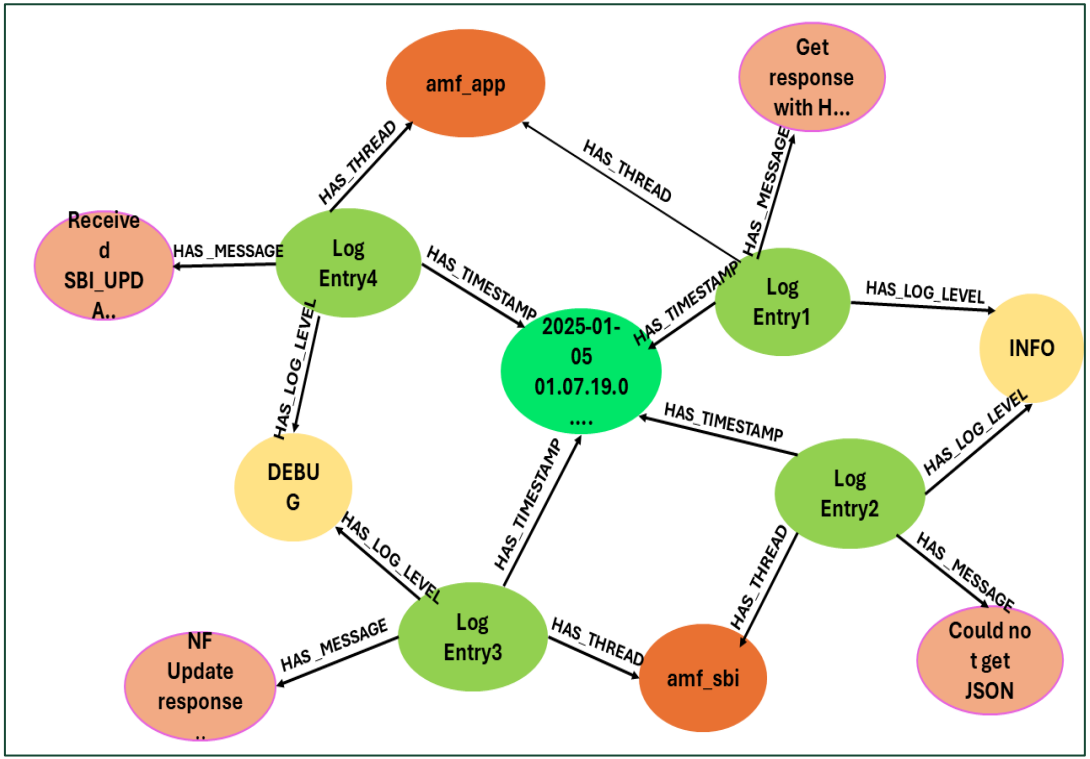
- 5G control-plane logs contain rich semantic identifiers that enable structured reasoning when transformed into a knowledge graph.

Semantic Transformation & Knowledge Graph

- Raw syslog data is ingested into the framework as input, capturing control-plane event information.
- Logs are parsed and normalized into structured fields such as timestamp, severity, component, and message content.
- The structured data is transformed into a knowledge graph (Neo4j) representing network behavior.
- The graph includes:
  - Nodes: Events, Timestamps, Network Functions (AMF, NRF, UPF), Message context
  - Edges: temporal relationships and semantic dependencies between events
- This transformation enables a unified, structured representation of heterogeneous log data.
- The knowledge graph encodes control-plane semantics, allowing efficient and meaningful querying.

Knowledge Graph

- Each log entry is mapped into a structured graph representation where nodes represent entities such as events, timestamps, severity levels, and network functions.
- Relationships encode temporal order and semantic dependencies, enabling explicit modeling of control-plane workflows.
- This representation allows efficient querying and structured reasoning across heterogeneous log data.



Knowledge Graph Data Efficiency

Reduced Query Data Size

- ❑ Knowledge graph retrieves only small, relevant subsets of data, avoiding full log scans.

Efficient Data Access

- ❑ Queries operate on indexed and structured data, reducing the effective data volume processed during analysis.

Minimized Runtime Data Transfer

- ❑ Instead of transferring large raw logs, only filtered event-level results are returned, lowering data movement.

Compression of Relevant Information

- ❑ The graph stores expanded structure, but queries extract compressed, meaningful insights from it.

Lower Processing Overhead per Query

- ❑ Even with higher storage, the system processes less data at runtime, improving efficiency.

AI-Driven Querying & Network Insights

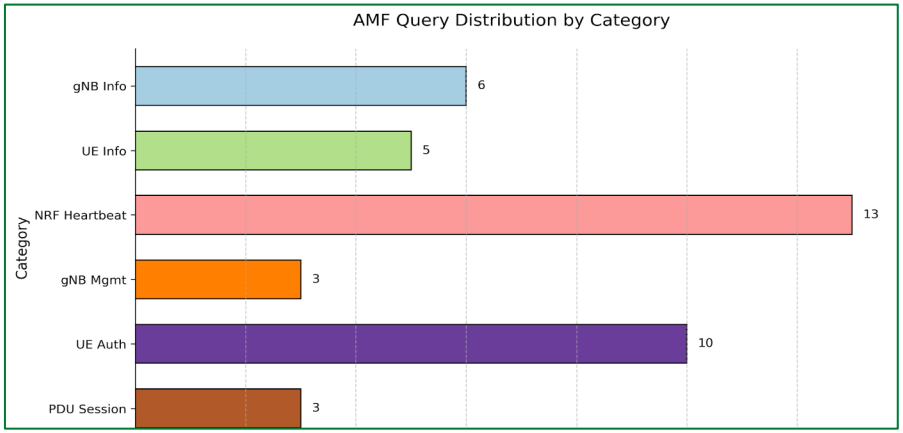
- An AI agent serves as the interface between the user and the knowledge graph.
- Users can issue natural language queries, eliminating the need for manual log analysis or query language expertise.
- The AI agent translates queries into Cypher queries, ensuring structured and deterministic retrieval.
- The LLM-powered query engine executes these queries on the knowledge graph to extract relevant events.
- The system produces:
  - Accurate event retrieval
  - Semantic analysis of network behavior
  - Anomaly detection and diagnosis
- This enables explainable, efficient, and scalable network management, beyond traditional log filtering or dashboard-based monitoring.

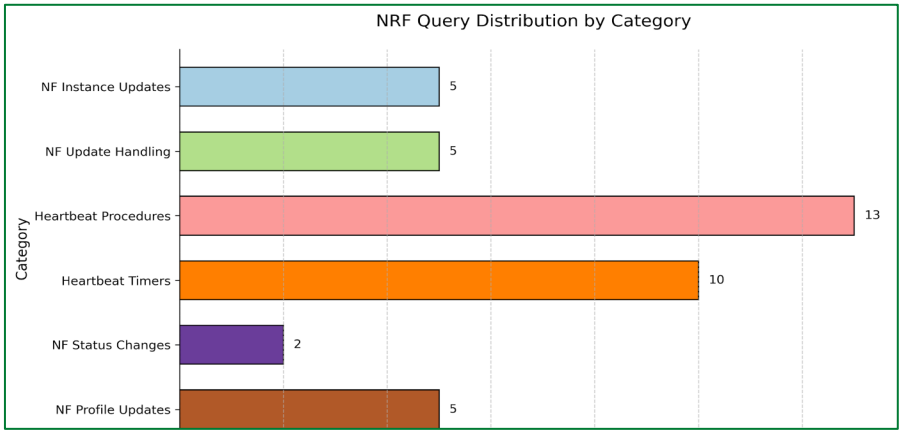
Crafting Queries based on Control Log planes – Manually and Code Iteration

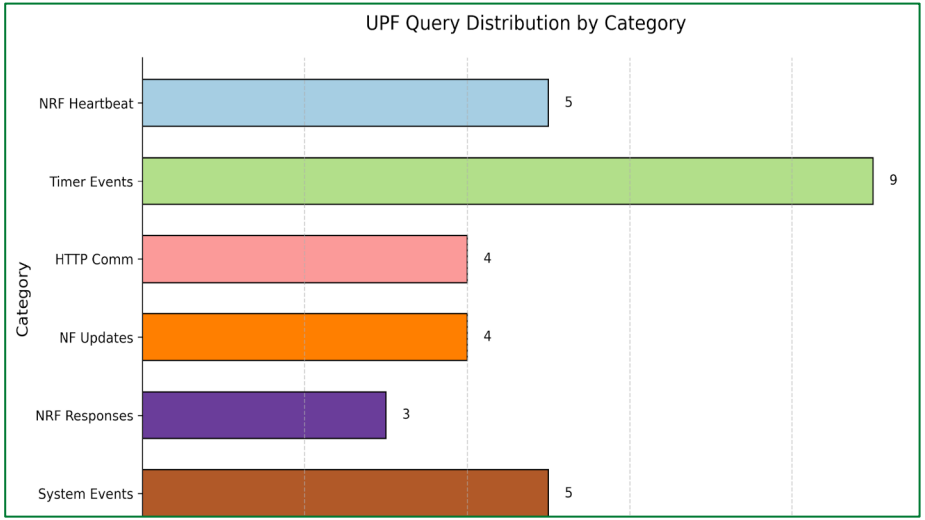
- We first analyzed AMF, NRF, and UPF logs manually to understand the structure of real 5G control-plane events.
- From this analysis, we identified important operational patterns, such as:
  - UE registration failures
  - NF discovery issues
  - Heartbeat losses
  - PDU session establishment events
  - Service status transitions
- Based on these patterns, we manually designed operator-style diagnostic queries that reflect realistic network management tasks.
- Each query was aligned with bounded event types and explicit semantic markers present in the logs.
- This manual process ensured that the queries were domain-specific, meaningful, and relevant to actual 5G operations.

- In addition to manual design, we developed a code-driven process to systematically analyze the dataset and support query generation.
- The pipeline iterates through AMF, NRF, and UPF logs to extract:
  - Recurring patterns
  - Event categories
  - Semantic markers such as IMSI, NF IDs, service names, and SBI transaction strings
- Pattern extraction is performed using regex-based parsing and keyword matching, aligned with log structure and message templates.
- The extracted patterns are mapped to event categories and operational scenarios, enabling automated generation of candidate queries.
- This approach ensures that queries are:
  - Data-driven (derived from actual logs)
  - Component-specific (AMF, NRF, UPF)
  - Semantically grounded in control-plane behavior
- The generated queries are then validated and refined manually to ensure correctness, completeness, and alignment with network management tasks.

Query distribution graphs for AMF, NRF, UPF







Enhancement Stages for Improving Accuracy

The system improves accuracy through a multi-stage enhancement pipeline, progressively refining how queries are generated and executed.

FGQ (Fine-Grained Query Decomposition):

- ❑ Breaks general queries into event-specific sub-queries
- ❑ Improves alignment with distinct control-plane activities (e.g., registration, session setup)

PSA (Prompting & Schema Alignment):

- ❑ Aligns the AI agent with the knowledge graph schema and log structure
- ❑ Ensures correct mapping of queries to nodes, relationships, and attributes

SPG (Semantic Pattern Grounding):

- ❑ Incorporates domain-specific patterns and keywords (e.g., IMSI, NF IDs, SBI messages)
- ❑ Enables consistent interpretation of log semantics across varying formats

LTQE (LLM-Tuned Query Execution):

- ❑ Enforces deterministic Cypher generation using low-temperature settings
- ❑ Includes validation/self-check mechanisms to ensure stable and accurate outputs

These staged enhancements transform the system from basic query generation to deterministic, semantically aligned, and highly reliable retrieval.

Evaluation Metrics for Retrieval Performance

The system is evaluated using four standard metrics to measure accuracy and semantic alignment of query results:

Precision

- ❑ Measures how many retrieved results are correct and relevant

Recall

- ❑ Measures how many relevant events are successfully retrieved
- ❑ High recall ensures important events are not missed

F1 Score

- ❑ Harmonic mean of precision and recall
- ❑ Reflects the balance between correctness and completeness

Jaccard Index

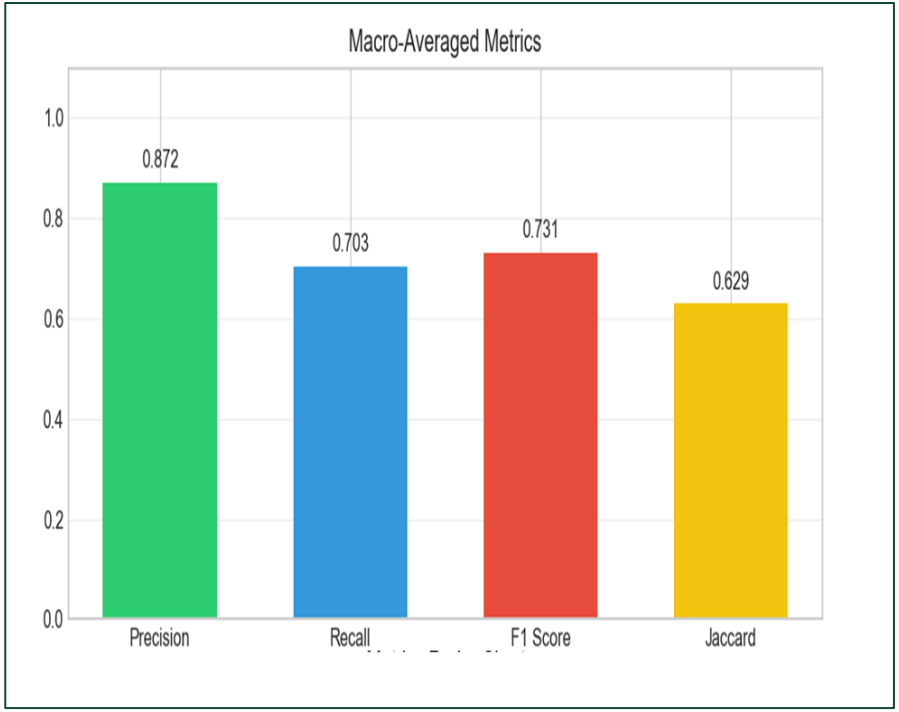
- ❑ Measures overlap between retrieved results and ground truth
- ❑ Provides a stricter evaluation of exact semantic matching

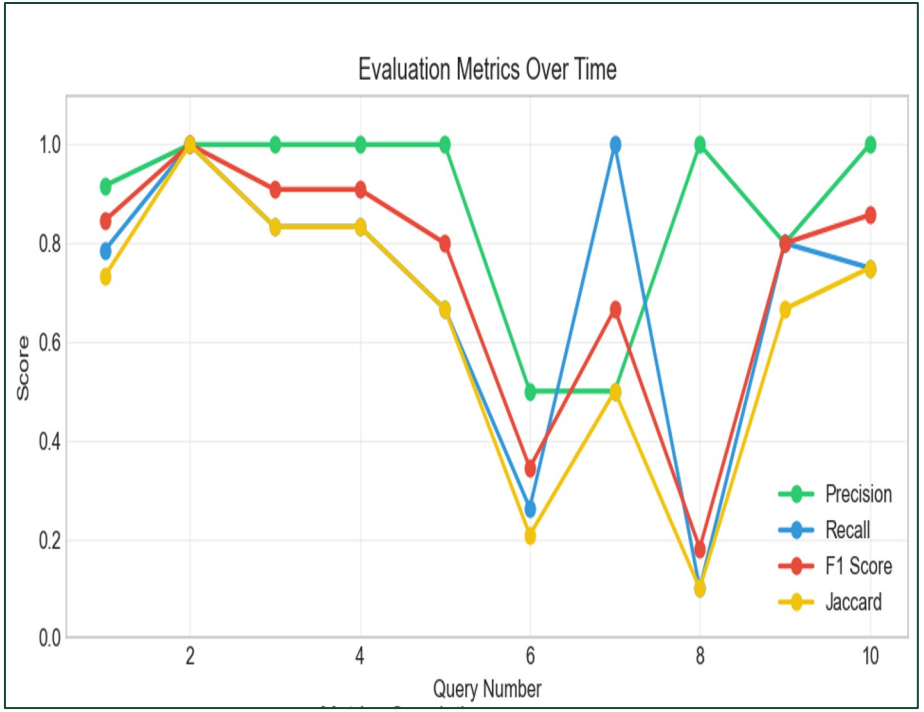
These metrics collectively evaluate the system's ability to perform:

- ❑ Accurate retrieval
- ❑ Complete event coverage
- ❑ Semantic consistency

Using multiple metrics ensures a comprehensive evaluation of both correctness and semantic alignment in query results.

Overall accuracy performance metrics before applying the Four Improvement stages





Macro-Averaged Metrics

- Before applying enhancements, the system achieves moderate performance with Precision (~0.87) higher than Recall (~0.70), F1 (~0.73), and Jaccard (~0.63).
- This indicates the model retrieves mostly correct results but misses relevant events and lacks full semantic alignment.

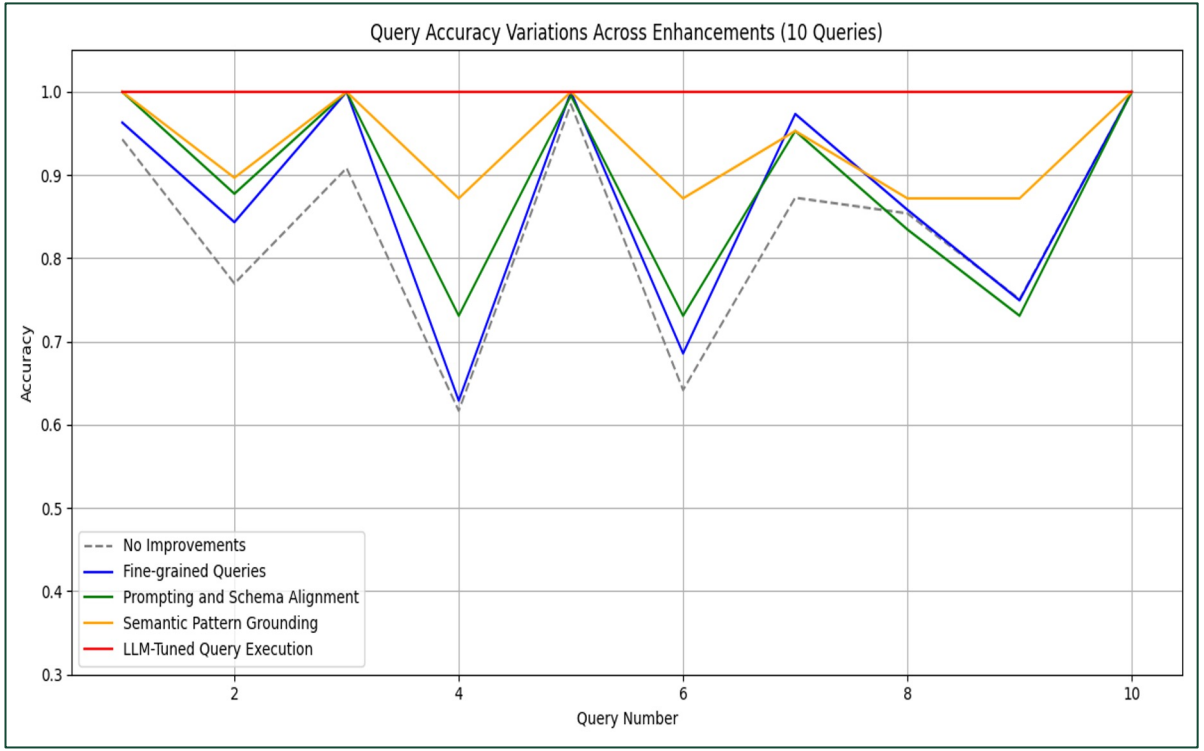
Evaluation Metrics Over Queries

- Performance varies significantly across queries, with some achieving near-perfect scores while others drop sharply, showing inconsistency and instability in baseline behavior.
- Lower scores in certain queries highlight ambiguity, incomplete understanding, and lack of semantic grounding before enhancements.

Query Accuracy Across Enhancements

Accuracy varies in early stages due to ambiguity but progressively stabilizes as enhancements are applied.

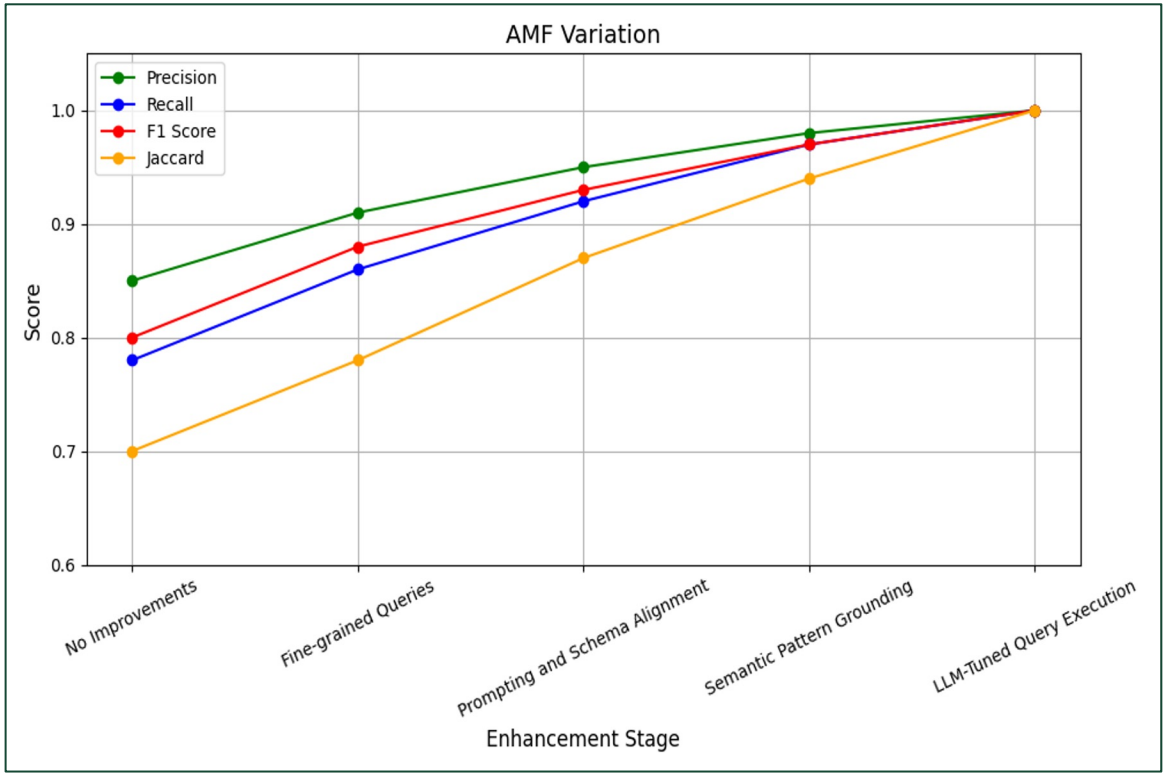
After LTQE, all queries converge to perfect accuracy, showing consistent, reliable, and deterministic query execution.



AMF Variation

All metrics (Precision, Recall, F1, Jaccard) steadily improve across enhancement stages, showing better alignment with control-plane events.

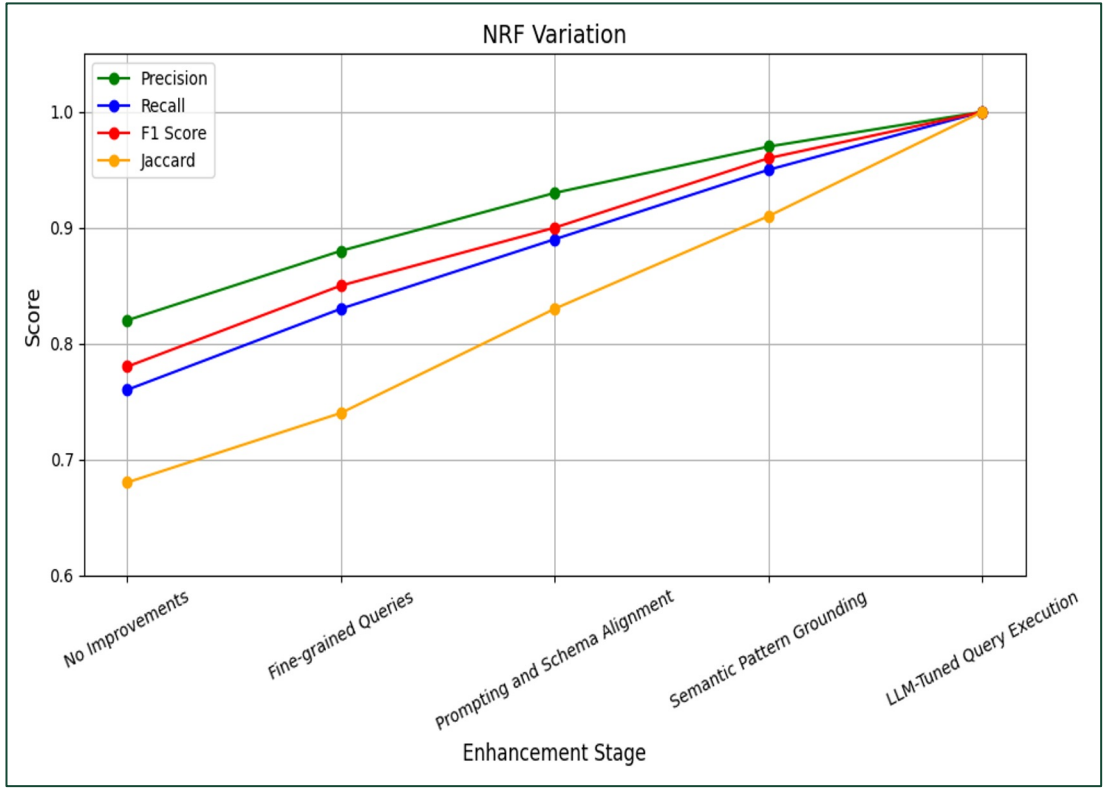
The final stage (LTQE) achieves perfect scores (1.0), indicating fully accurate and deterministic query generation for AMF logs



NRF Variation

Like AMF, NRF shows consistent improvement across all metrics, with gradual reduction in mismatch and ambiguity.

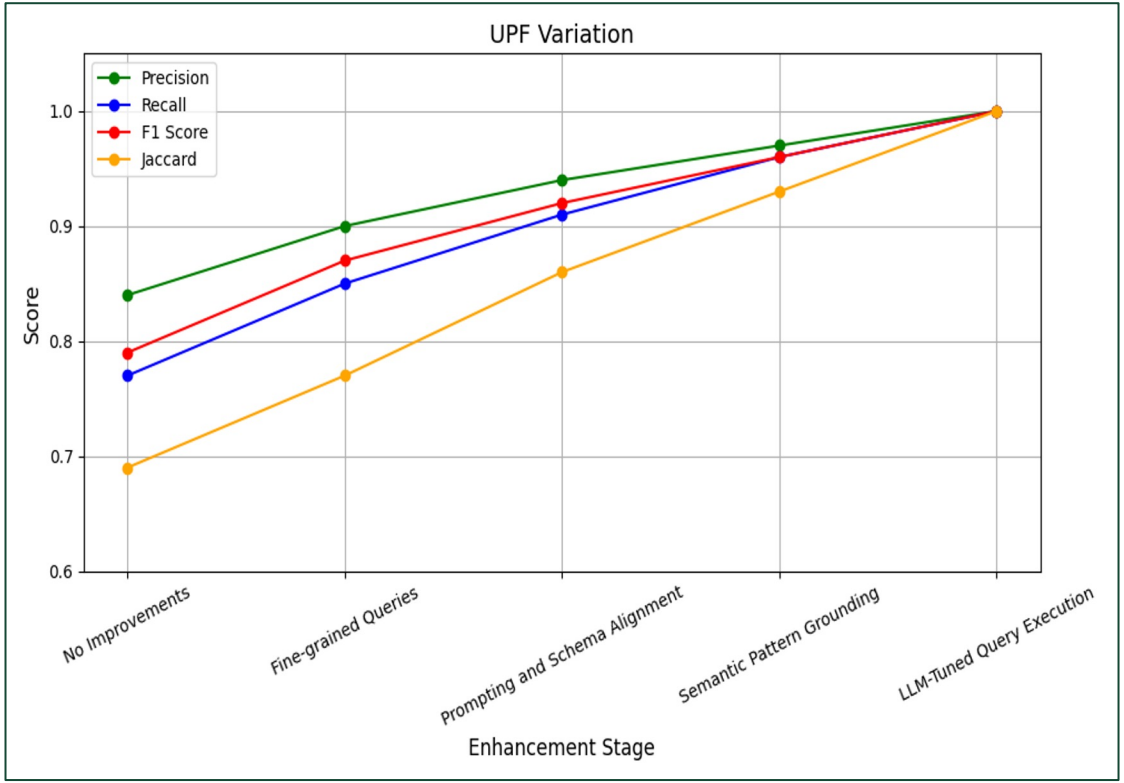
The enhancements enable accurate semantic interpretation of NRF-specific events like NF registration and discovery.



UPF Variation

UPF follows the same trend, with noticeable improvement especially after semantic grounding stages.

Final convergence at 1.0 confirms precise handling of session-level and data-path events in UPF logs.



Log Data Reduction and Efficiency

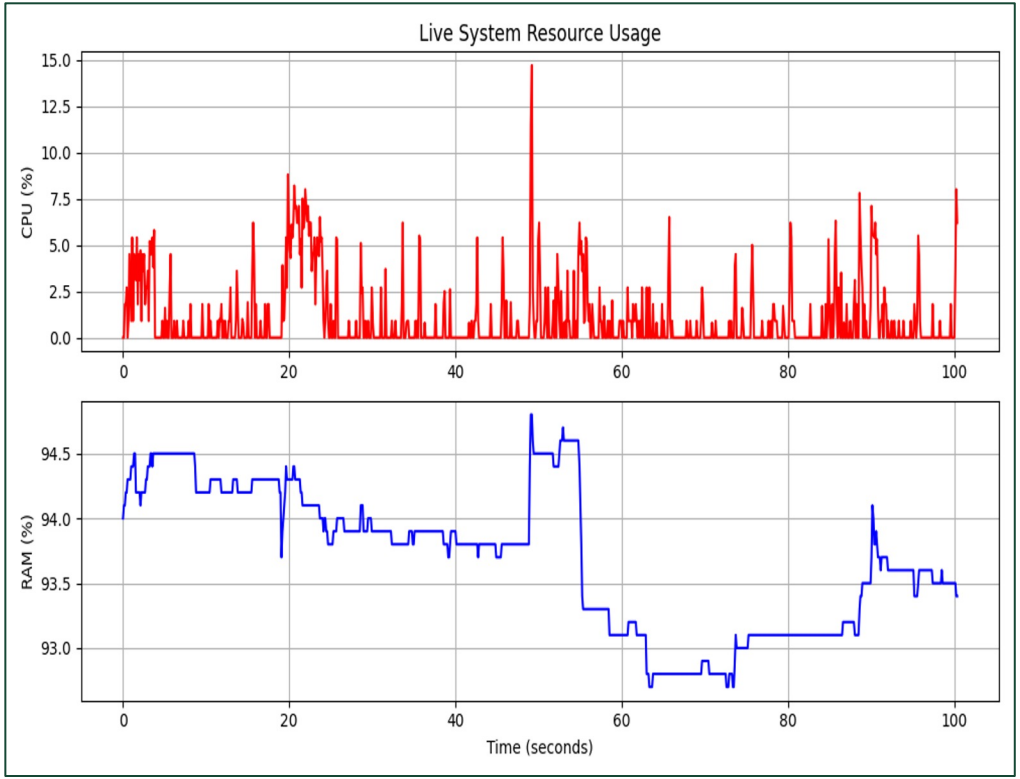
Query Theme	Raw Log Size Queried (Lines)	Filtered Log (Lines)	% Reduction
All UPF messages from NRF	1000	125	87.5%
NRF Heartbeat Communication	1000	375	62.5%
PDU Session Establishment	1000	125	87.5%
UE Context Release	1000	1	99.9%

- The framework significantly reduces the volume of data processed during queries by retrieving only relevant events. For example, certain queries achieve up to 99.9% reduction in log volume compared to raw log scanning.
- This reduction improves system efficiency, reduces computational overhead, and enables faster response times.
- By focusing on event-specific retrieval, Aim5B enhances operational scalability.

System Performance

The framework demonstrates low resource usage, with peak CPU utilization around 16% and memory usage around 2%.

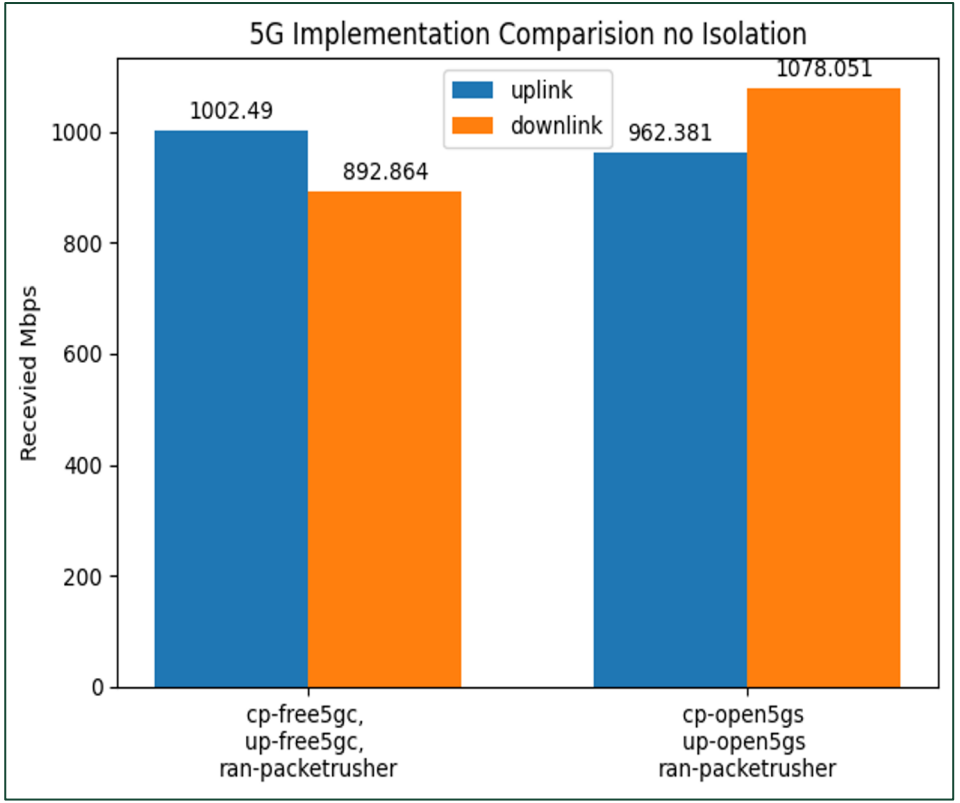
This indicates that the system is lightweight and scalable for deployment in real network environments.



Throughput Analysis & Relevance to AIM5B

- The comparison demonstrates that 5G core implementations (free5GC and Open5GS) exhibit different throughput characteristics, with performance varying across uplink and downlink directions.
- Under baseline conditions (no isolation), one system achieves higher uplink throughput, while the other shows stronger downlink performance, reflecting differences in internal processing and resource handling.
- When isolation (resource separation) is introduced, the impact is implementation-dependent:
  - Some systems benefit from reduced contention and show performance gains
  - Others experience reduced effective throughput, particularly in received data, despite similar transmission rates
- Connection to AIM5B:
  - These results highlight that overall system performance is sensitive to processing overhead and resource utilization.
  - AIM5B improves efficiency by reducing control-plane overhead, transforming raw logs into a knowledge graph and enabling targeted, event-level querying.
  - Instead of scanning large volumes of logs, the system retrieves only relevant semantic information, minimizing unnecessary computation.

Stack	Uplink Gain with Isolation	Downlink Gain with Isolation
free5gc	1003 → 1058 Mbps (small gain)	896 → 980 Mbps (noticeable gain)
open5gs	1847 → 1778 Mbps (small drop)	1999 → 1999 Mbps (same send, but recv dropped from 1078 to 879 Mbps)



Conclusion

- Aim5B provides a scalable and interpretable approach to network management by combining semantic modeling with AI-driven querying. It enables accurate and efficient analysis of control-plane events and reduces operational complexity.
- The framework demonstrates the potential of combining knowledge graphs and Artificial Intelligence to address the challenges of modern network observability.
- The system is validated on a private 5G testbed, demonstrating high accuracy, efficiency, and scalability.

Future Work

- Future work will extend the framework to support multi-modal reasoning by integrating additional data sources such as metrics, topology, and configuration into the knowledge graph.
- This extension will enable cross-layer diagnosis and deeper understanding of network behavior, moving toward AI-native network management for 6G systems.

Questions?

Thank You