



LogsLLM: Automated Log Analysis and Management with Large Language Models

GÉANT Incubator Programme, WP6

Dimitris Pantazatos, NTUA

16th SIG-NGN Meeting in Collaboration with SIG-AI, Karlsruhe, Germany

10 October 2025

Public (PU)

Agenda

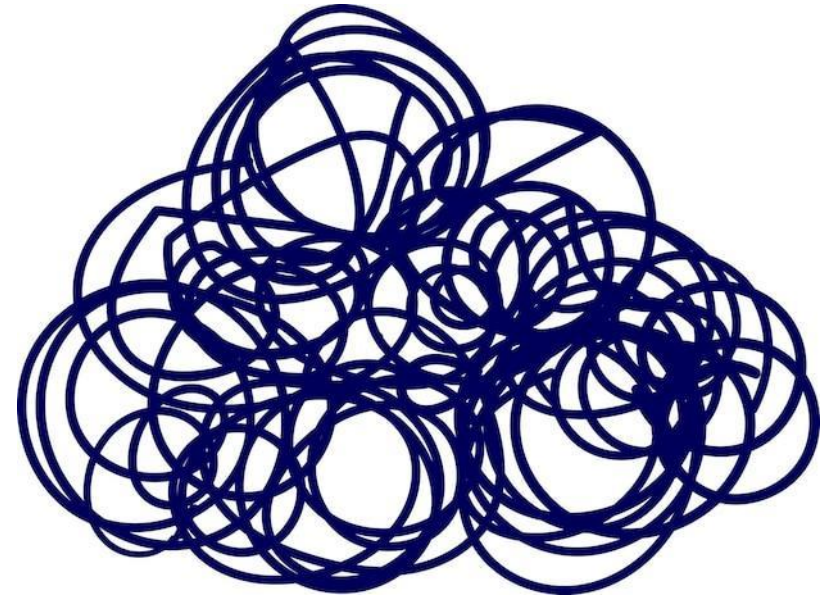
- Background & Problem
- What LogsLLM Does
- How LogsLLM Helps NOCs and NRENs
- Indicative Use Cases
- Case Study
- System Architecture and Implementation
- Evaluation Results
- Conclusion and Future Work

Partners



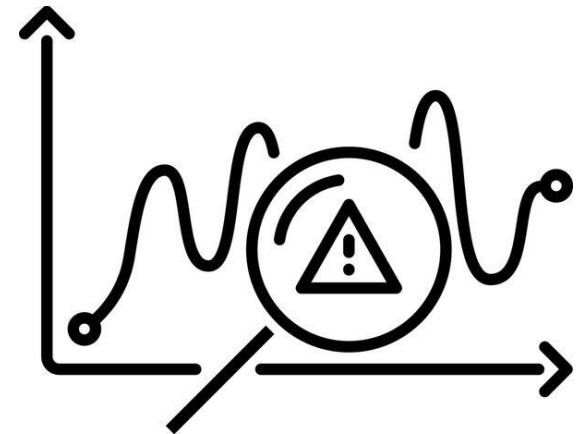
Background & Problem Statement

- NRENs & NOCs deal with massive heterogeneous logs.
- Logs are:
 - Unstructured, inconsistent formats.
 - Hard to normalize and correlate.
 - Require expert interpretation.
- Traditional tools: rule-based, rigid, limited adaptability.



What LogsLLM Does

- Standardizes heterogeneous logs into a unified, structured format
- Detects anomalies and correlates related events across systems
- Generates insights in natural language to support NOC operators
- Outcome:
 - Faster incident understanding
 - Reduced manual analysis
 - Consistent, explainable log intelligence



How LogsLLM Helps NOCs and NRENs



Comparative Positioning

- Compared to Splunk, ELK, Graylog, Datadog:
 - Strengths: flexibility, schema-first, explainability, openness.
 - Limitations: scale, latency, fewer built-in enterprise features.
- Positioning: complementary layer (adds explainability).

Indicative Use Cases

Security Monitoring

- Scenario: Firewall shows surge in failed SSH logins.
 - Logs standardized (event_type=failed_login, severity=critical).
 - Summary: “2,500 failed attempts from 203.0.113.x in 10 minutes.”
 - Operator escalates as brute-force attempt.
- Benefit: Reduced chance of missed alerts.

Cross-System Correlation

- Scenario: Firewall blocks port traffic + Icinga outage.
 - Correlated output: “Outage coincides with firewall blocking port 443 from subnet A.”
 - Operator investigates misconfiguration or malicious block.

Training/Knowledge Transfer

- Junior NOC engineers see LogsLLM summaries alongside raw logs.
- Summaries help them learn log meaning faster.
- Reduces onboarding time.

Case Study – Event Volume Analysis

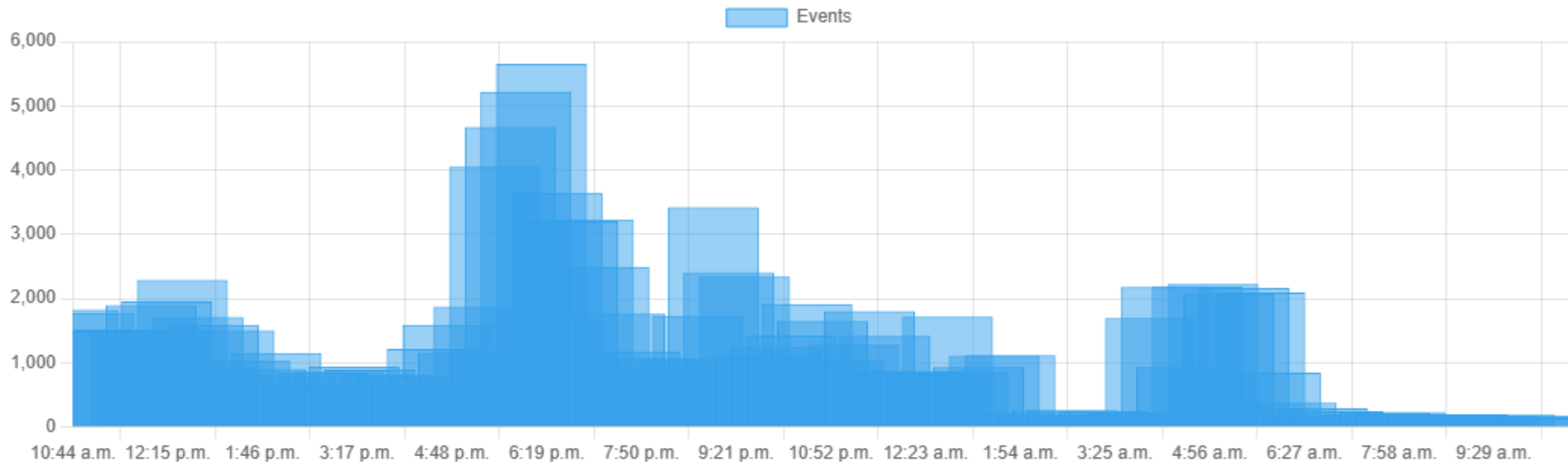
From (UTC) 21/03/2025 10:44 am

To (UTC) 22/03/2025 10:44 am

IP (e.g., 203.0.113.7) CIDR (e.g., 203.0.113.0/24) src or dst

Pattern (substring or regex) Regex Services CSV (e.g., icinga,fire) 200 **Search**

Timeline



There was a high volume of events (126327) across the specified time window, with peak activity occurring around 18:14 on March 21st. Next query: "Top users" to identify which specific users or devices were responsible for the majority of these events.

Case Study - Device & IP Attribution

LogsLLM — Chat ← Back to Dashboard

FILTERS

From (local)	To (local)	Services
<input type="text" value="21/03/2025 10:44 am"/>	<input type="text" value="22/03/2025 10:44 am"/>	<input type="text" value="firewall, uptimokuma"/>
IP	CIDR	Role
<input type="text" value="203.0.113.7"/>	<input type="text" value="203.0.113.0/24"/>	<input type="text" value="src or dst"/>
Pattern	Regex	Limit (rows to ground)
<input type="text" value="substring or regex"/>	<input type="text" value="No"/>	<input type="text" value="10000"/>

Tip: Leave filters empty to use plain chat. "Ask With Filters" grounds the answer on exactly the matched rows.

Can you find the top devices or IPs for this time range?

Based on the provided rows digest, I can extract some information about the top devices and IPs.

****Top Devices (Hosts):****

- * 'null' appears 7736 times
- * '200.136.95.235' appears 528 times
- * 'fw01' appears 4 times

****Top Source IPs:****

- * '52.247.164.25' appears 1103 times
- * '218.153.149.249' appears 910 times
- * '31.137.147.224' appears 859 times
- * '179.122.59.99' appears 850 times

Examples: "Critical anomalies on 2025-03-20?", "Did firewall rejects coincide with Icinga downtime yesterday?", "Top NetBox changes by user last week".

Case Study - Malicious Activity Assessment

LogsLLM — Chat
← Back to Dashboard

FILTERS

From (local)	To (local)	Services
21/03/2025 10:44 am	22/03/2025 10:44 am	firewall, uptimekuma
IP	CIDR	Role
203.0.113.7	203.0.113.0/24	src or dst
Pattern	Regex	Limit (rows to ground)
substring or regex	No	10000

Tip: Leave filters empty to use plain chat. "Ask With Filters" grounds the answer on exactly the matched rows.

can some of them to be considered malicious?

Based on the provided rows digest, I can analyze the events for potential malicious activity.

There are 5218 'connection_blocked' events and 4663 'connection_denied' events in the top events list. However, without further context or specific criteria for what constitutes "malicious," it's difficult to determine if these events are indeed malicious.

Some of the blocked connections appear to be incoming (Direction: in) with various protocols (TCP, UDP), but I don't see any evidence of known malicious patterns or indicators of compromise (IOCs). The 'Reason' field for most of these events is "match," which could indicate a rule-based block rather than an actual security threat.

To determine if some of these connections can be considered malicious, more information about the rules and configurations used to generate these events would be

Examples: "Critical anomalies on 2025-03-20?", "Did firewall rejects coincide with Icinga downtime yesterday?", "Top NetBox changes by user last week".

Ask about anomalies, trends, correlations...

Ask

High-Level Architecture

- Data ingestion layer: firewall, Icinga, NetBox, Uptime Kuma.
- Agent 1: Standardization: transforms raw logs → unified schema.
- Agent 2: Analysis: detects anomalies, correlations, summaries.
- Core LLM engine: prompt-driven reasoning.
- Storage: PostgreSQL, JSON schema.
- APIs: FastAPI endpoints

Multi-Agent Design

- Agent 1: Standardization
 - Input: raw, heterogeneous logs.
 - Output: structured JSON (timestamp, severity, message).
 - Uses prompt-driven parsing.
- Agent 2: Analysis
 - Input: normalized logs.
 - Output: anomalies, correlations, summaries.
 - Uses LLM-guided contextual reasoning.

Evaluation of the Standardization Agent - Overview

- Goal: assess how well the first agent converts heterogeneous logs into a unified schema.
- Dimensions:
 - Detection – can it recognize and classify events?
 - Accuracy – are fields correct and consistent?
 - Utility – are results informative for operators?
- Verdict levels: PASS / WARN / FAIL

Evaluation of the Standardization Agent - Datasets & Models

- Datasets: Icinga | NetBox | Uptime Kuma | Firewall
- Models tested:
 - ChatGPT 4.1-mini
 - DeepSeek R1-14B
 - LLaMA 3.1-8B
- All models used the same schema & prompts.

Evaluation of the Standardization Agent - Key Results by Dataset

Model	Detection (Avg)	Accuracy (Avg)	Utility (Avg)	Overall Comment
ChatGPT 4.1-mini	0.95	0.98	0.61	Highest detection & accuracy — strong, stable results
DeepSeek R1-14B	0.60	0.90	0.58	Rich enrichment (utility) but weaker detection
LLaMA 3.1-8B	0.80	0.98	0.52	Reliable & schema-compliant outputs

Evaluation of the Analysis Agent - Evaluation Overview

- Setup:
 - 4 representative cases (spikes, cross-source correlation, device/IP attribution, config impact).
 - Models tested: ChatGPT 4.1-mini, LLaMA 3.1-8B, DeepSeek R1-14B.
 - 10 runs / model / case with identical parameters (Temp 0.2, Max Tokens 8000).
- Metrics:
 - Objective:
 - Claims Accuracy ($\pm 10\%$),
 - Entity Match (Jaccard ≥ 0.5),
 - Severity Match,
 - Inter-run Stability.
 - Human Evaluation (0-5): Usefulness | Clarity | Trust.

Evaluation of the Analysis Agent - Key Results

Model	Claims Accuracy	Entity Match (Jaccard)	Severity Match	Usefulness (0–5)	Clarity (0–5)	Trust (0–5)	Verdict Summary
ChatGPT 4.1-mini	0.96	0.85	1.00	4.55)	4.3	4.9	Highest precision and interpretability
LLaMA 3.1-8B	0.72	0.65	1.00	3.85	4.0	4.0	Reliable reasoning; moderate contextual depth
DeepSeek R1-14B	0.67	0.58	1.00	3.0	4.0	4.0	Readable outputs; limited analytical accuracy

Conclusion

1

**LogsLLM =
Proof-of-
concept for
LLMs in log
management**

2

**Multi-agent
approach
shows
strong
potential**

3

**Complements
existing
monitoring
tools**

4

**Sets the
stage for
future AI-
powered
NOC
operations**

Future Work

1

**Instruction
-tuned
Q&A over
logs**

2

**RLHF for
improved
accuracy**

3

**Real-time
alerting
integration**

4

**Broader
deployment
across
GEANT &
nmaas**



Thank You

www.geant.org



Co-funded by
the European Union