Trustworthy Zero-touch Network and Service Management in 6G Networks with XAI and LLMs

Abdelkader Mekrache, Akram Boutouchent, Adlen Ksentini, Christos Verikoukis

Co-funded by the European Union

6G-INTENSE (Intent-driven NaTive AI architecturE supporting Compute-Network abstraction and Sensing at the Deep Edge) proposes a new system architecture for future 6G Smart Networks, characterised by high performance and energy efficiency, facilitating advanced internet applications. Key goals include driving an industry revolution, fostering digital transformation, and building smart societies with improved quality of life through features, like autonomous systems, haptic communication, and smart healthcare.

1. PoC Context & Overview

Zero-touch Network and Service Management (ZSM) enables autonomous network management without human intervention. ZSM includes detecting anomalies, identifying the root causes of anomalies, and autonomously resolving these anomalies.

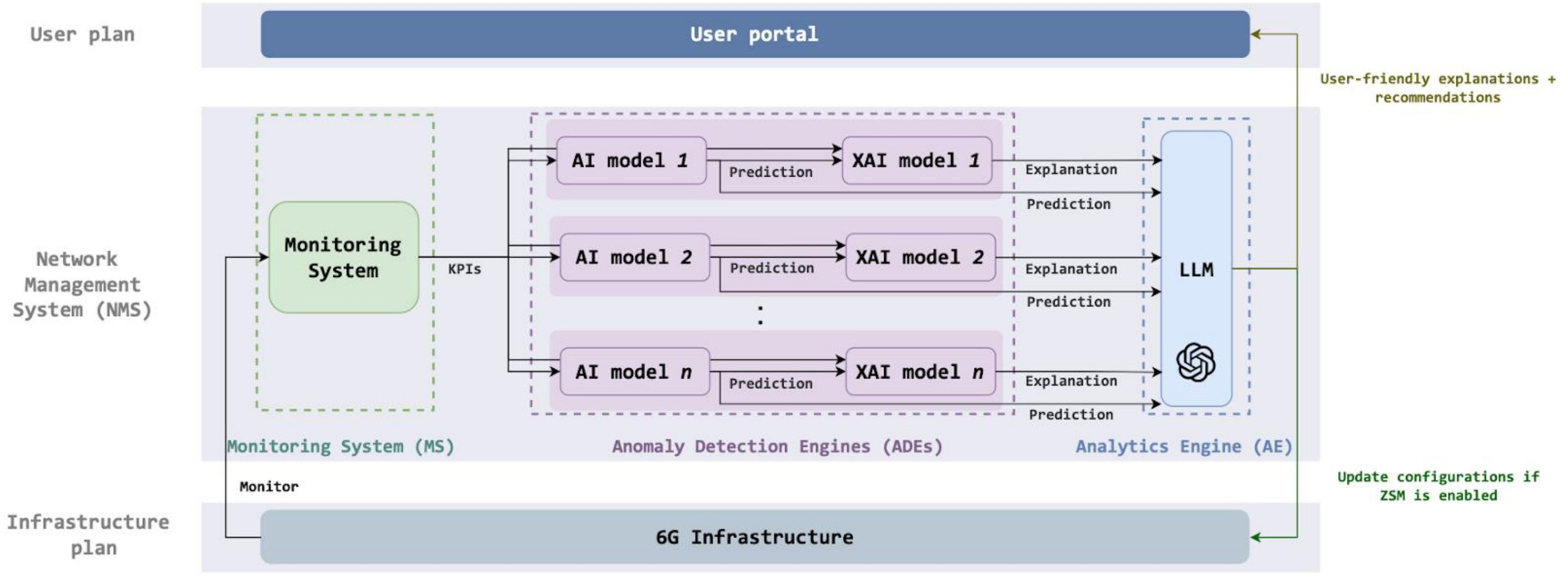
Currently, AI methods have been widely used in research to detect anomalies; however, these methods are often black boxes lacking explainability, making it difficult to extract the root cause.

XAI methods have emerged to explain AI decisions, thus identify the root causes of anomalies. However, these methods often rely on numerical values to explain anomalies, which can be difficult for users with little domain knowledge to understand => Trust ✘ !

Using natural language to explain the anomalies => LLMs are a promising choice to enable trustworthy ZSM.

In addition, LLMs excel at reasoning tasks, making them a candidate for anomaly resolution based on their explanations.

2. PoC Architecture



LLM-enabled trustworthy ZSM architecture design [1].

The proposed ZSM framework features a closed-control loop for managing 6G services, is illustrated. It consists of three layers:

- □ **6G Infrastructure:** includes cloud/edge clusters and radio units supporting 6G services.
- □ **NMS:** autonomously detects and resolves anomalies in the infrastructure.
- □ **User Plane:** involves users deploying 6G services and interacting with the NMS. They receive human-like explanations from the NMS regarding detected anomalies, recommendations to resolve them, or actions taken by the NMS if ZSM is enabled.

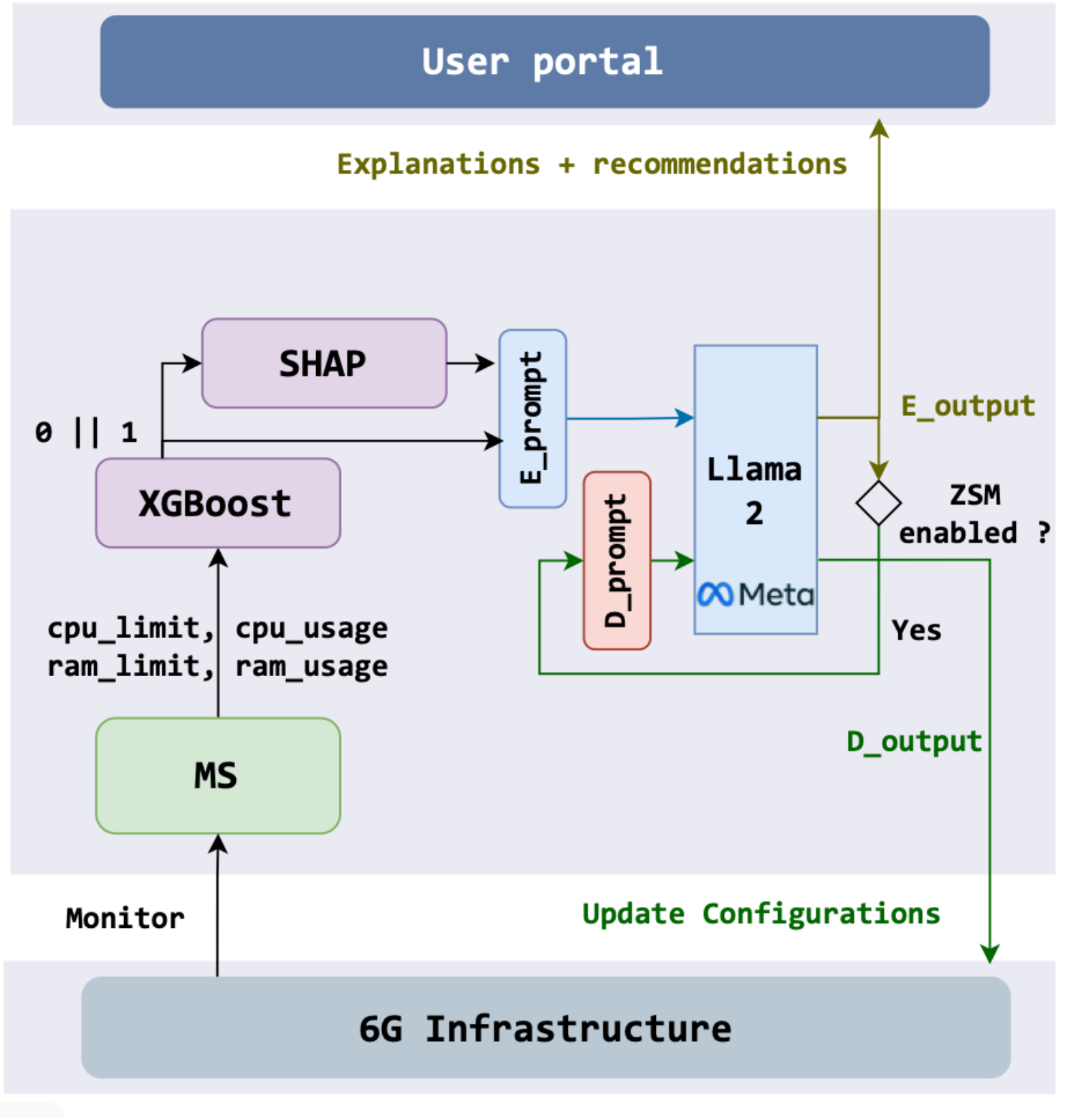
The design of the NMS consists of three stages:

- □ **Monitoring System (MS):** collects KPIs from the 6G infrastructure.
- □ **Anomaly Detection Engines (ADEs):** use AI to detect anomalies and XAI to provide numerical explanations.
- □ **Analytics Engine (AE):** provides comprehensive explanations for root causes and suggests actions.

3. PoC Demonstration

The demonstrated PoC ensures that CPU and RAM resources for a given microservice application are dynamically adjusted to prevent SLA latency violations. For this purpose, we employed:

- □ **XGBoost** as the AI model, which predicts whether the microservice application will violate SLA latency constraints.
- □ **SHAP** for XAI, which identifies the root cause of a potential violation, whether it is due to CPU, RAM, or both.
- □ **Llama2** as the LLM, which explains to users why an SLA violation was predicted and how to resolve it.



Demonstration use case [1].

4. PoC Experimentation Setup

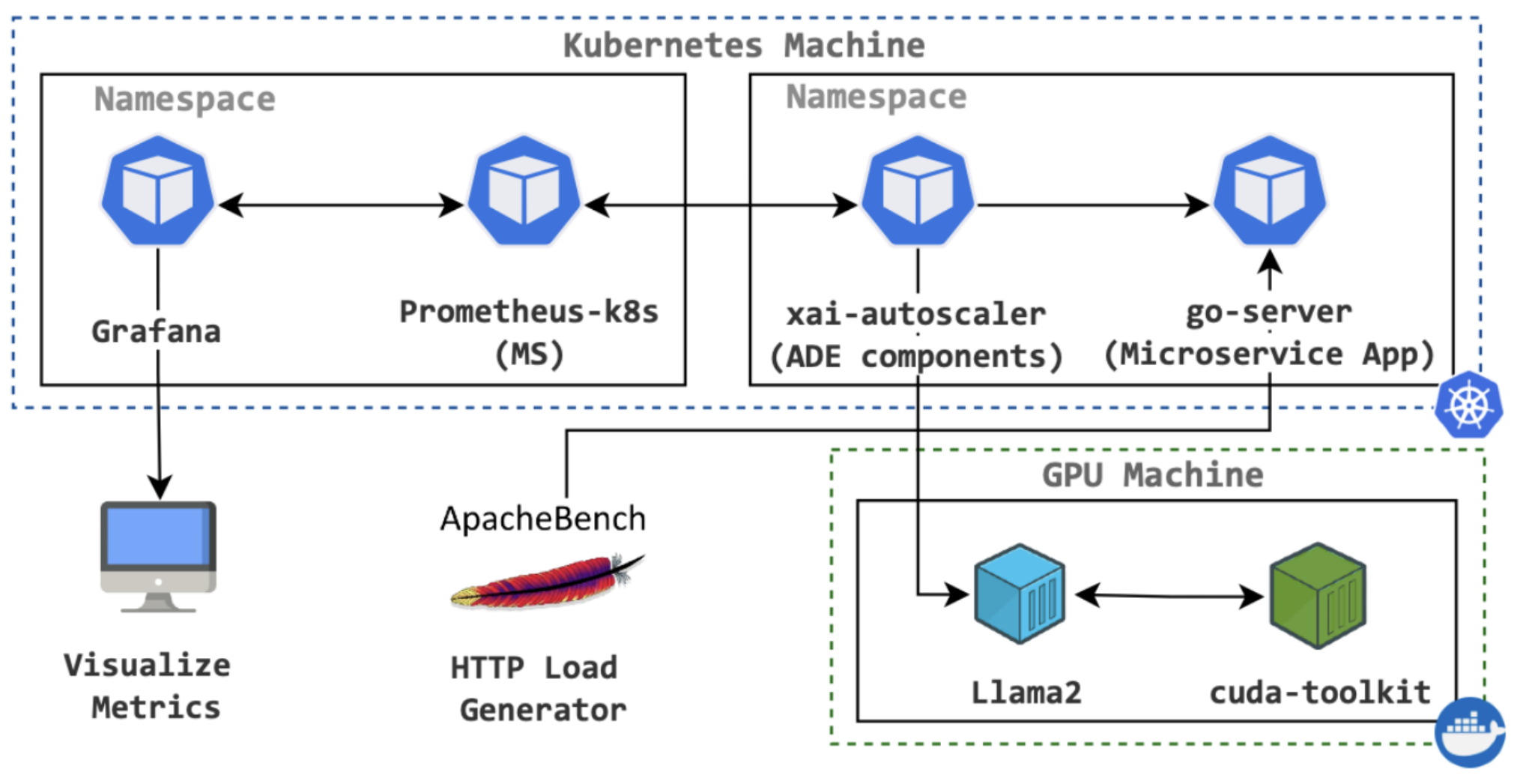
Two machines are involved:

Machine 1 (Cluster Host):

- □ Runs Kubernetes managing a single-node cluster.
- □ Hosts Microservices: (MS) and Anomaly Detection Engine (ADE) components.

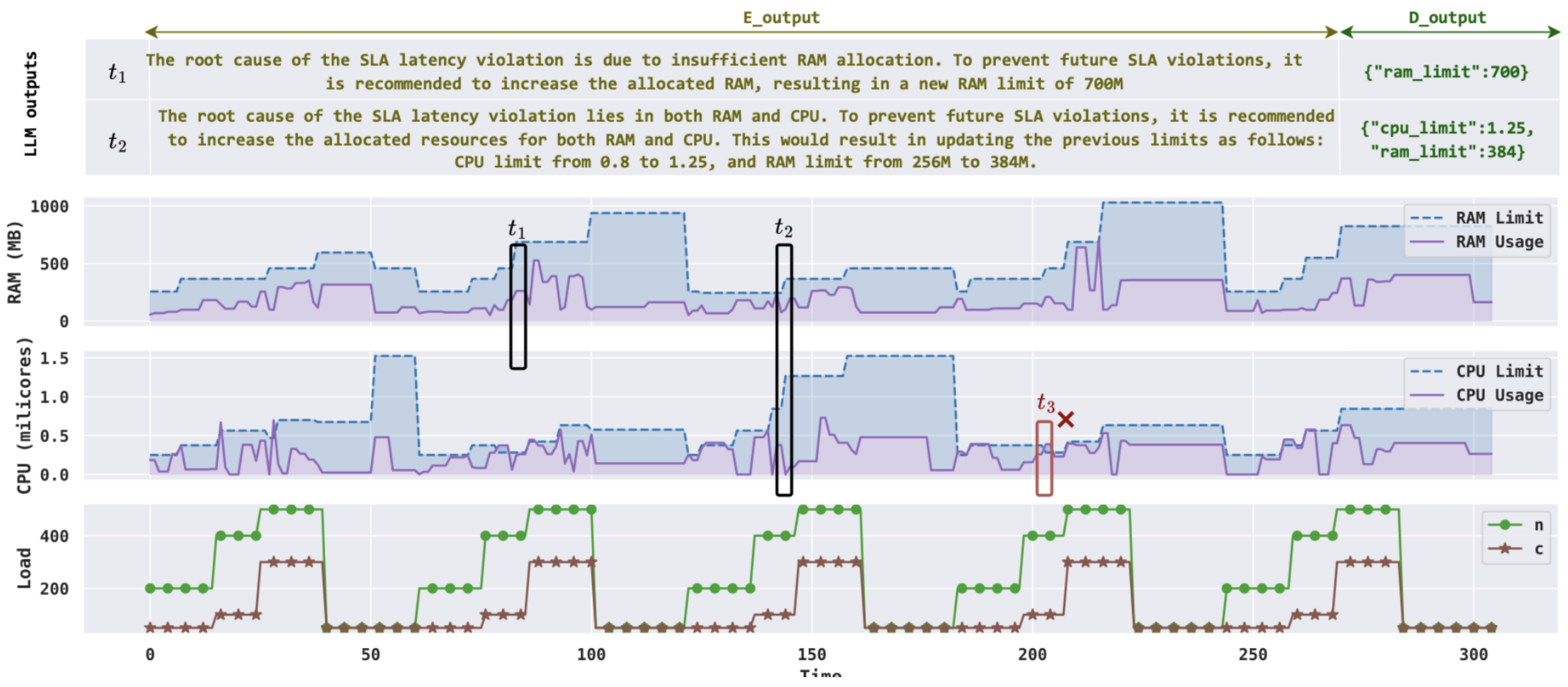
Machine 2 (LLM Host):

- □ NVIDIA A100 GPU with 40GB vRAM
- □ Runs Llama2 LLM via textgen-webui in Docker



Experimentation setup.

5. PoC Experimentation Results



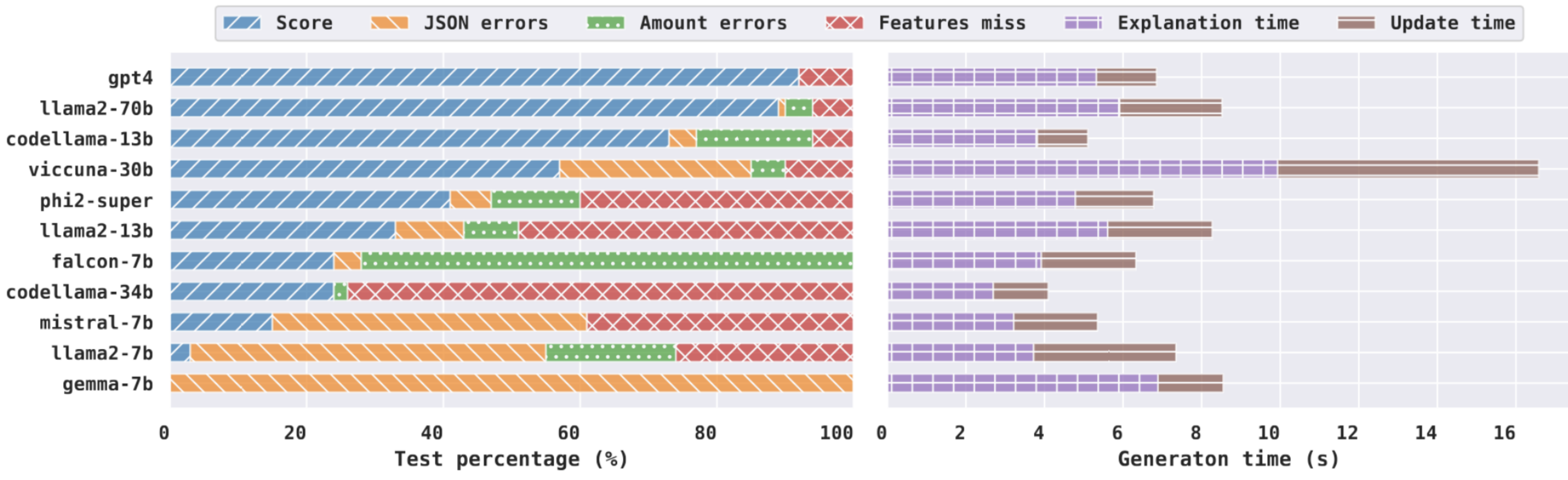
Dynamic CPU and RAM scaling in response to microservice load [1].

Scenario:

- Microservice app initiated with 0.25 CPU cores and 256 MB RAM
- Load generated using ApacheBench with varied request patterns
- Monitoring via Prometheus, collecting CPU, RAM, and LLM outputs

Results:

- The approach dynamically adjusted CPU/RAM to match load
- Llama2 generated E_output and D_output
- Minor CPU allocation errors observed at specific timestamps (e.g., t_3)



LLMs scores and E_output-D_output generation times for the dymanic scaling use case [1].

Scenario:

- Generated 100 random CPU/RAM allocations with Random SHAP values: –5 to 5
- Evaluated LLMs (open-source & GPT-4) with scoring: +1 if output was correct (structure, relevant features, increased values); 0 otherwise

Results:

- GPT-4 scored 91/100, best overall.
- Best open-source: Llama2 70B (high accuracy)
- Best trade-off: CodeLlama 13B (accuracy vs. speed)

6. Conclusion

- □ Introduction of a trustworthy ZSM framework for 6G networks.
- □ A framework that combines AI (XGBoost) for anomaly detection, XAI (SHAP) for root cause analysis, and LLMs (Llama2/GPT-4) for explanation & corrective actions.
- □ Evaluation results achieved accurate SLA violation prediction and dynamic resource updates.
- □ Human-readable explanations, boosting user trust and transparency.

[1] Mekrache, Abdelkader, et al. "On Combining XAI and LLMs for Trustworthy Zero-Touch Network and Service Management in 6G." IEEE Communications Magazine (2024).