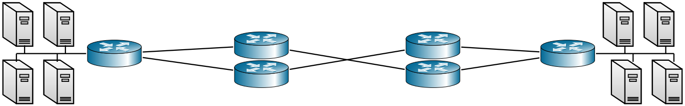
Incident Response Planning Using a Lightweight Large Language Model with Reduced Hallucination

Network and Distributed System Security (NDSS) Symposium
San Diego, USA, February 24, 2026

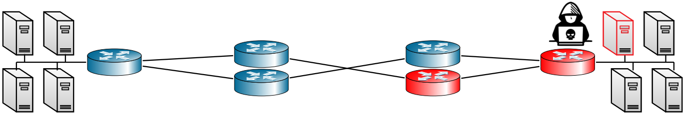
Kim Hammar, Tansu Alpcan, and Emil Lupu
kim.hammar@unimelb.edu.au

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE

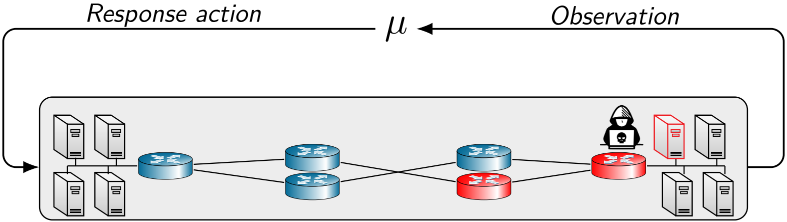
Problem: Incident Response



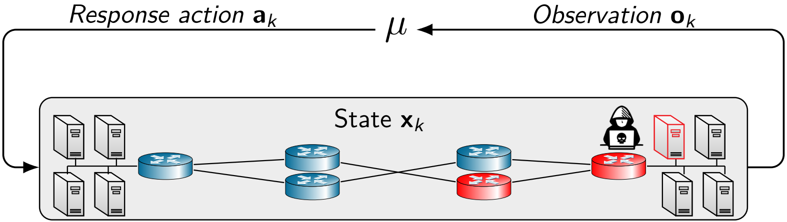
Problem: Incident Response



Problem: Incident Response



Problem: Incident Response



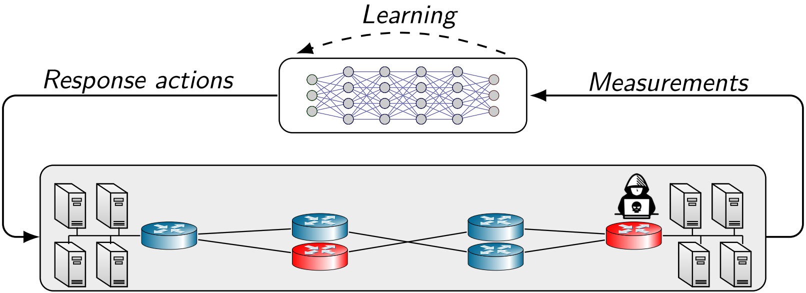
- Hidden state $\mathbf{x}_k$ (e.g., which components are compromised?).
- Observation $\mathbf{o}_k$ (e.g., log files and security alerts).
- Response action $\mathbf{a}_k$ (e.g., update network segmentation).
- **Goal**: find a strategy μ that meets response objectives.

The Traditional Approach to Incident Response

- Incident response is **managed by security experts**.
- We have a global shortage of more than 4 million experts.
- Pressing need for new decision support systems!
  - Current approach: *response playbooks*.

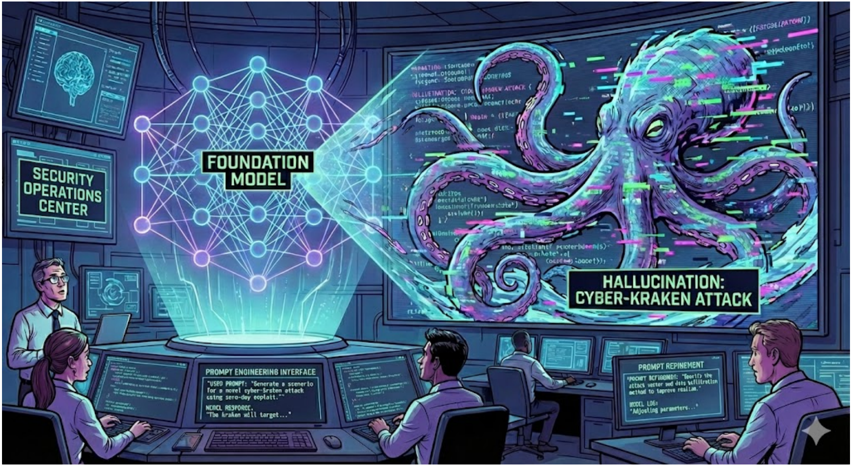
Next Generation Incident Response System

- Promising approach: **Leverage LLMs for decision support.**
- In 2025:
  - IBM launched Instana, an agentic incident response system.
  - Google launched CodeMender, an AI agent for code security.
  - Microsoft launched security copilot, an AI agent for security.

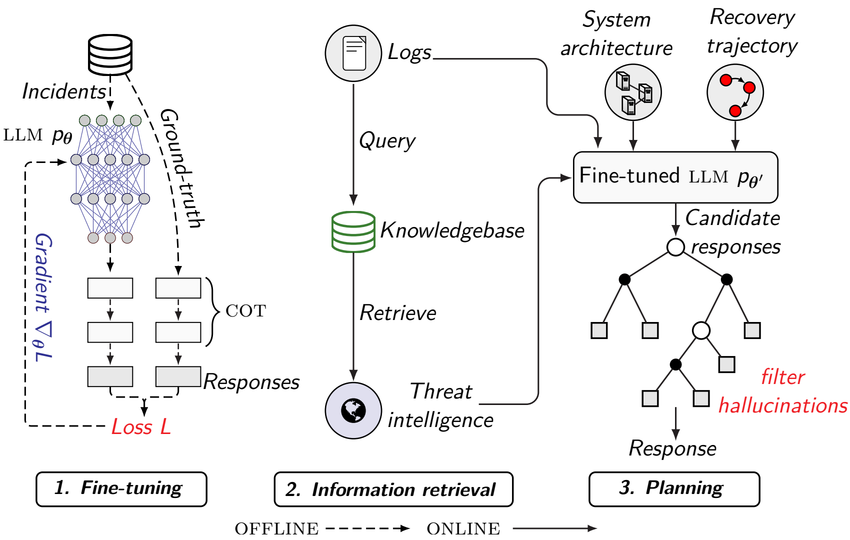


Limitations of Current Systems

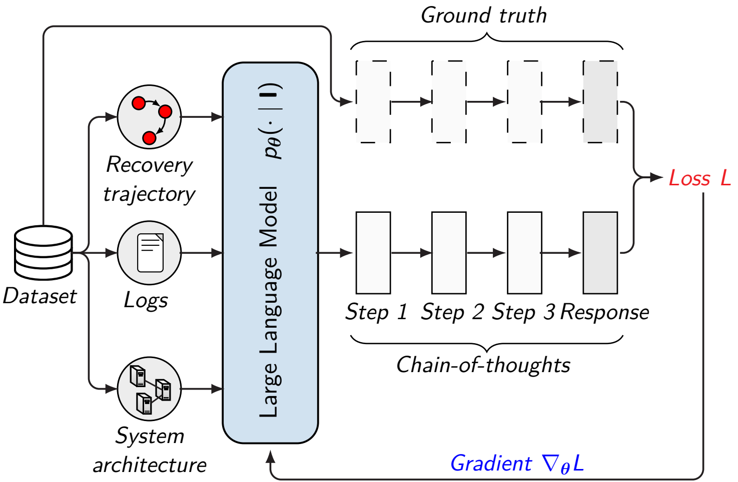
- Current systems have the following limitations:
  - They rely on an external LLM provider.
  - They have no theoretical guarantees.
  - They are prone to **hallucinations**.



Our Method for Incident Response Planning

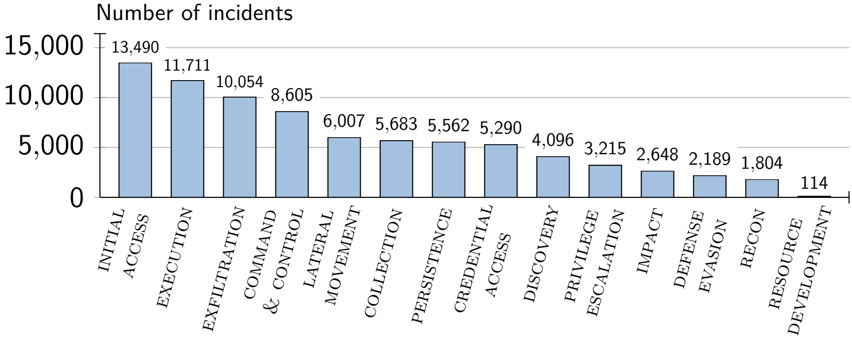


Fine Tuning a Lightweight LLM



- **Supervised training** based on a dataset of 68,000 incidents.
- Cross-entropy loss function.

Constructing the **Fine-Tuning Dataset**



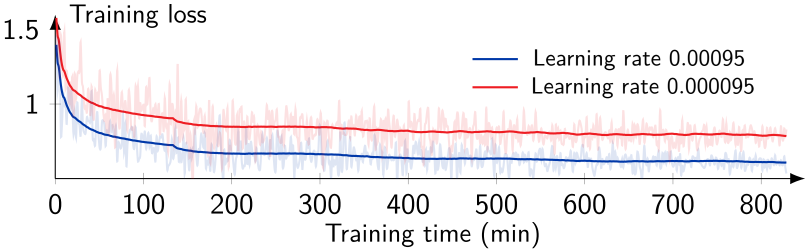
- We use a **combination of testbed data and synthetic data.**
- We start by emulating 500 incidents on our testbed.
- We then prompt foundation models to generate new incidents.
- Our dataset is available on huggingface.

Instruction Fine-Tuning

- We fine-tune the DEEPSEEK-R1-14B LLM on a dataset of 68,000 incidents $\mathbf{x}$ and responses $\mathbf{y}$.
- Minimize the **cross-entropy loss:**

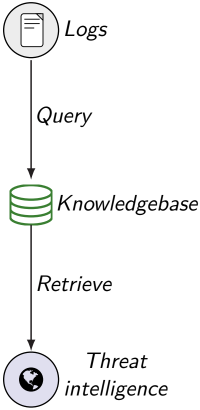
$$L = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{m_i} \ln p_{\theta}\left(\mathbf{y}_k^i \mid \mathbf{x}^i, \mathbf{y}_1^i, \ldots, \mathbf{y}_{k-1}^i\right),$$

where m_i is the length of the vector $\mathbf{y}^i$.

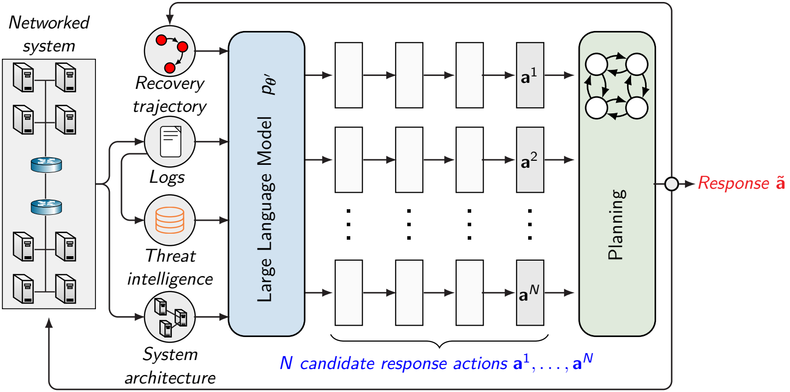


Retrieval-Augmented Generation (RAG)

- We use regular expressions to extract indicators of compromise (IOC) from logs.
  - e.g., IP addresses, vulnerability identifiers, etc.
- We use the IOCs to **retrieve information about the incident** from public threat intelligence APIs, e.g., OTX.
- We include the retrieved information in the context of the LLM.



Using the LLM for **Incident Response Planning**



- Instead of selecting the first action generated by the LLM, we
  - **Generate** N ***candidate actions***.
  - Evaluate each candidate through lookahead optimization.

Hallucinated Response Action

Definition 1 (informal)

A response action $\mathbf{a}_t$ is hallucinated if it does not make any progress towards recovering from the incident.

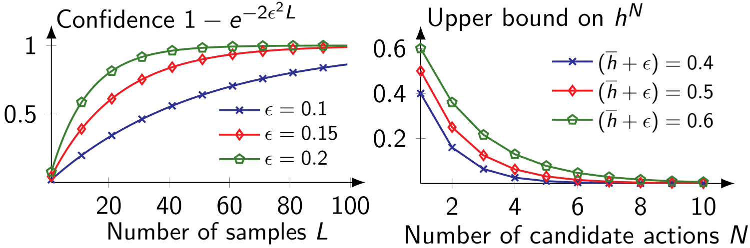
Chernoff Bound on the Hallucination Probability

Proposition 1 (Informal)

- *Let h be the true hallucination probability.*
- *Let $\overline{h}$ be the empirical probability based on L samples.*

We have

$$P(h \geq \overline{h} + \epsilon) \leq e^{-2\epsilon^2 L}.$$



Conditions for **Lookahead to Filter Hallucinations**

Proposition 2 (Informal)

- *Let η be the total variation between LLM's predictions and true system dynamics.*
- *Let δ be the minimal **difference in recovery time between a hallucinated and non-hallucinated action.***
- *Assume at least one candidate action is not hallucinated.*

If

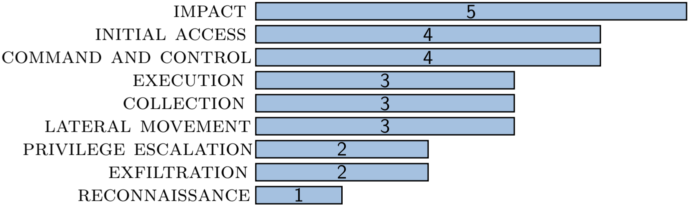
$$\delta > 2\eta \|J\|_\infty \left(\|\tilde{J}\|_\infty + 1 \right),$$

then the selected action will not be hallucinated.

Experimental Evaluation

- We evaluate our system on 4 public datasets.

Dataset	*System*	*Attacks*
CTU-Malware-2014	Windows xp sp2 servers	Various malwares and ransomwares.
CIC-IDS-2017	Windows and Linux servers	Denial-of-service, web attacks, SQL injection, etc.
AIT-IDS-V2-2022	Linux and Windows servers	Multi-stage attack with reconnaissance, cracking, and escalation.
CSLE-IDS-2024	Linux servers	SambaCry, Shellshock, exploit of CVE-2015-1427, etc.



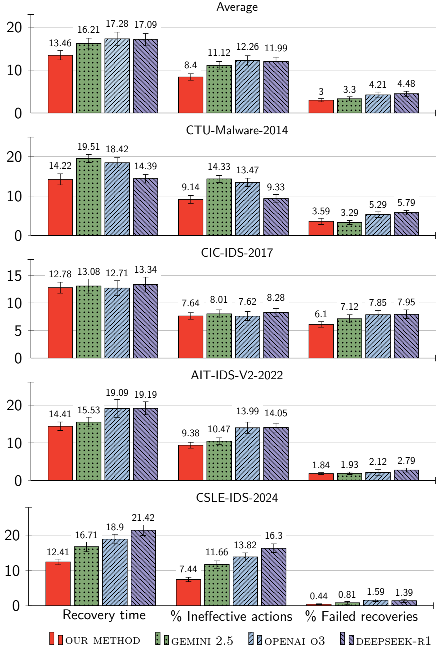
Distribution of MITRE ATT&CK tactics in the evaluation datasets.

Baselines

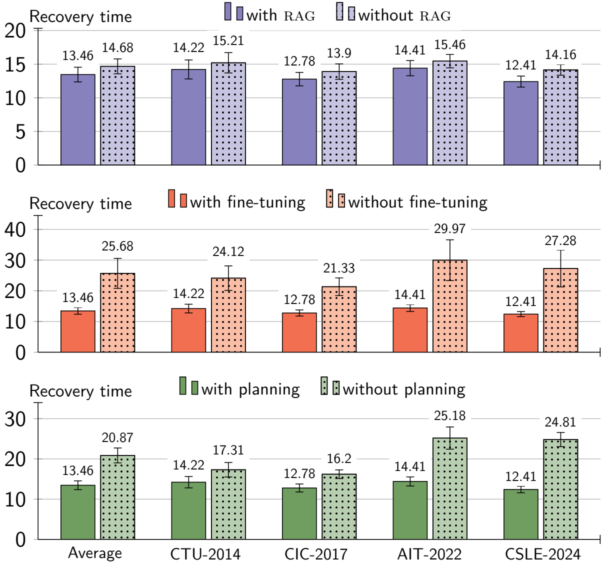
- We compare our system against **frontier LLMs.**
- Compared to the frontier models, our system is lightweight.

System	*Number of parameters*	*Context window size*
OUR SYSTEM	14 billion	$128,000$
DEEPSEEK-R1	671 billion	$128,000$
GEMINI 2.5 PRO	unknown ($\geq$ 100 billion)	1 million
OPENAI O3	unknown ($\geq$ 100 billion)	$200,000$

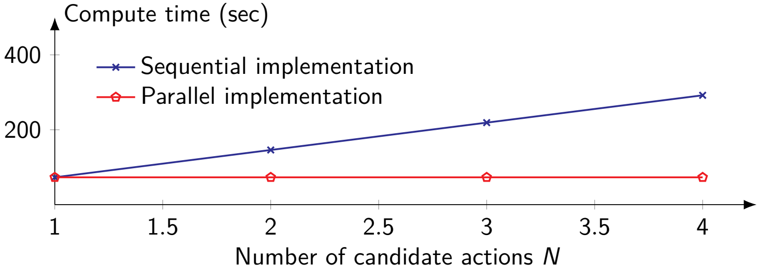
Evaluation Results



Ablation Study



Scalability



- The lookahead optimization is computationally intensive since it requires making multiple inferences with the LLM.
- The computation can be parallelized across multiple GPUs.

Conclusion

- **LLMs will play a role as decision support in SOCs.**
  - Remarkable knowledge management capabilities.
  - Hallucination remains a challenge.

- We present a **lightweight** method for response planning.
  - Allows to control the hallucination probability.
  - Significantly outperforms frontier LLMs.
  - Demo: https://www.youtube.com/watch?v=SCxq2ye-R4Y.