Incident Response Planning with a Foundation Model

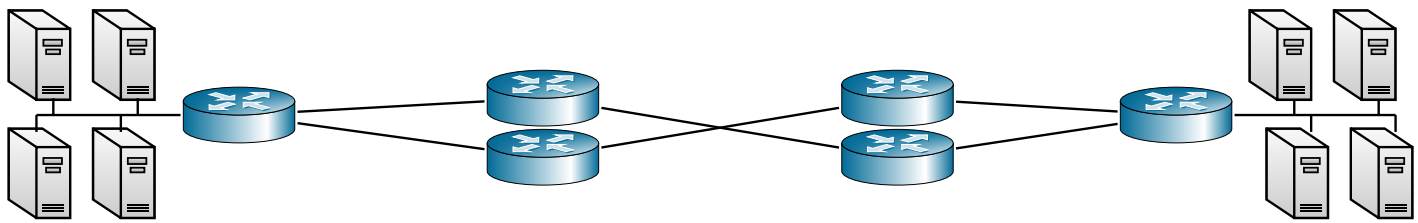
University of Melbourne
December 5, 2025

Dr. Kim Hammar
kim.hammar@unimelb.edu.au

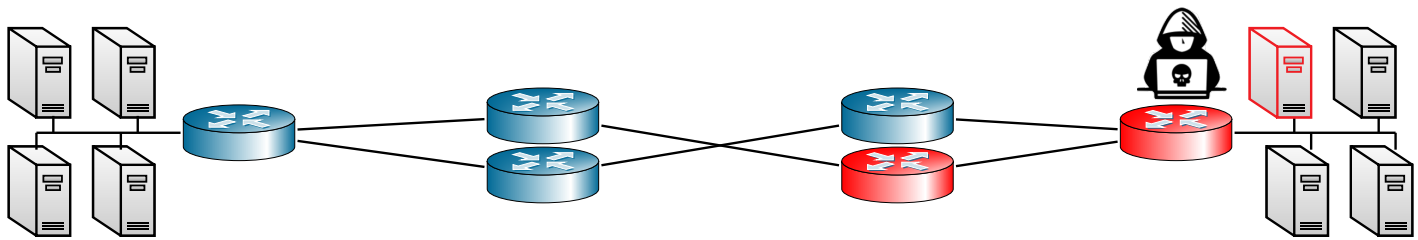
Paper: *Incident Response Planning Using a Lightweight Large Language Model with Reduced Hallucination*
(*Kim Hammar, Tansu Alpcan, and Emil Lupu*)

Accepted to NDSS Symposium 2026
Preprint: https://arxiv.org/abs/2508.05188

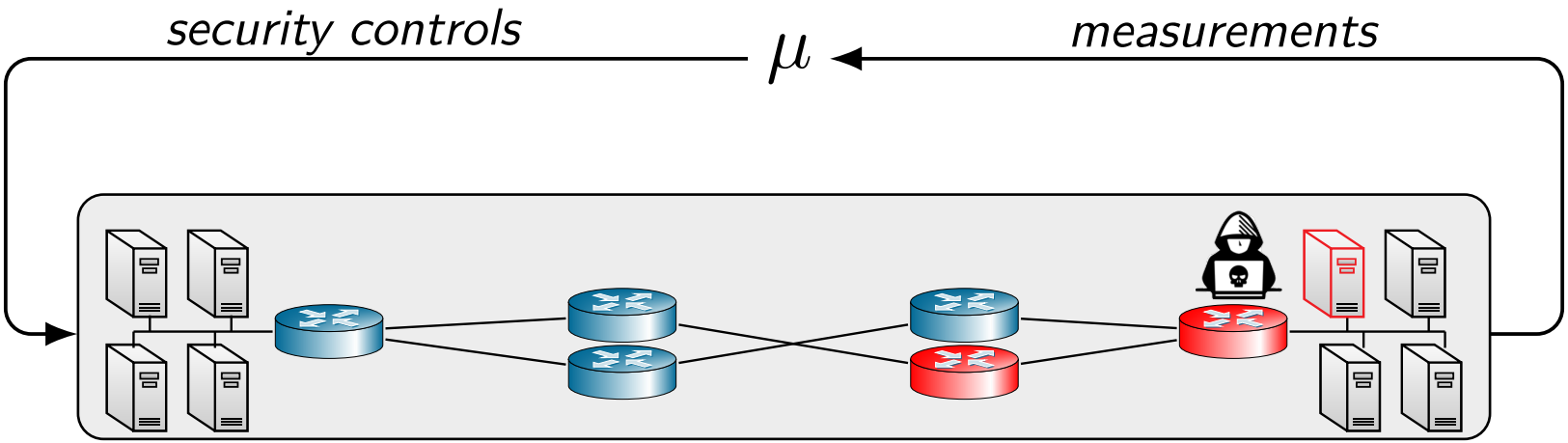
Problem: Incident Response



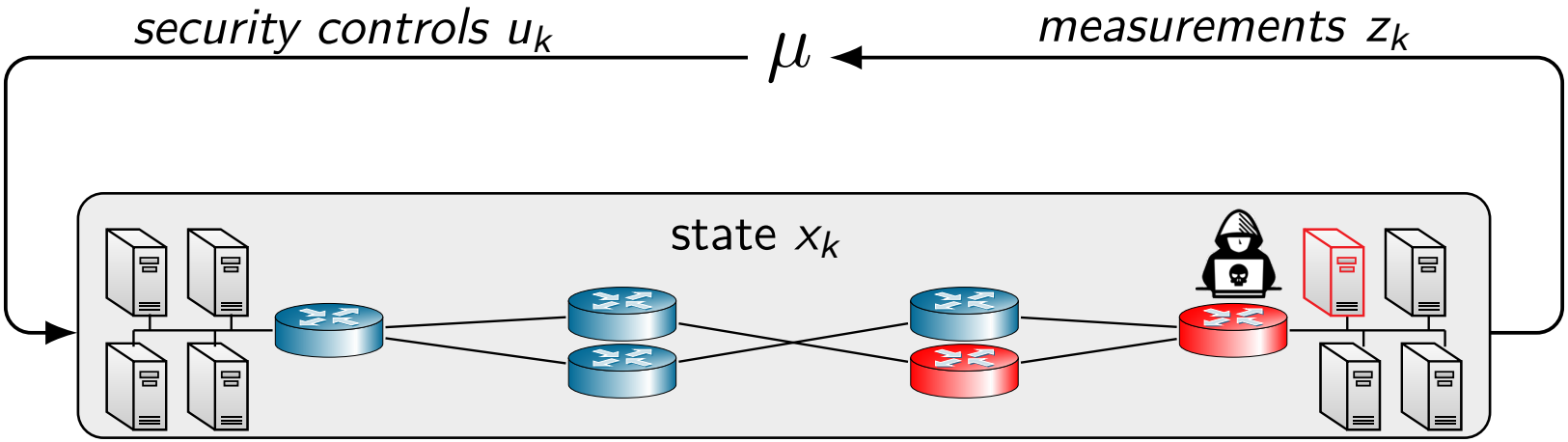
Problem: Incident Response



Problem: Incident Response



Problem: Incident Response

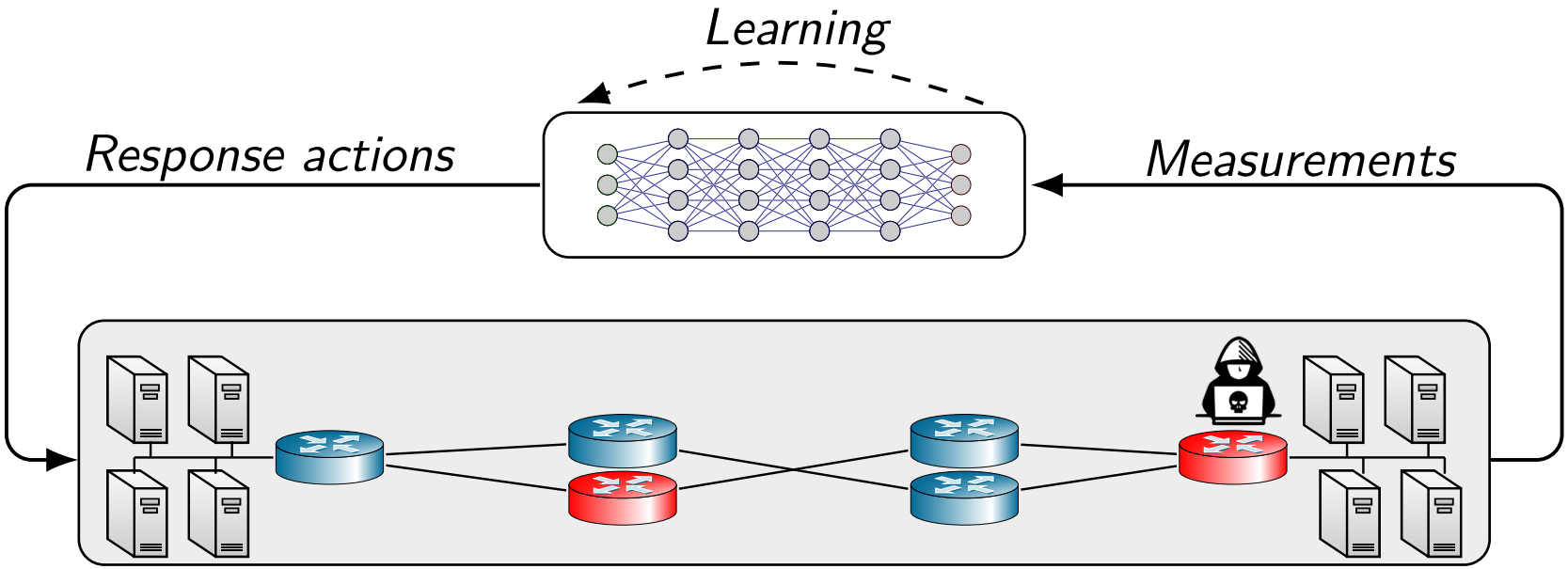


- Hidden states x_k, transition probabilities $p_{ij}(u)$.
- Observation z_k generated with probability $p(z_k \mid x_k, u_{k-1})$.
- Control u_k.
- **Goal**: find a policy μ that meets response objectives.

Current Practice

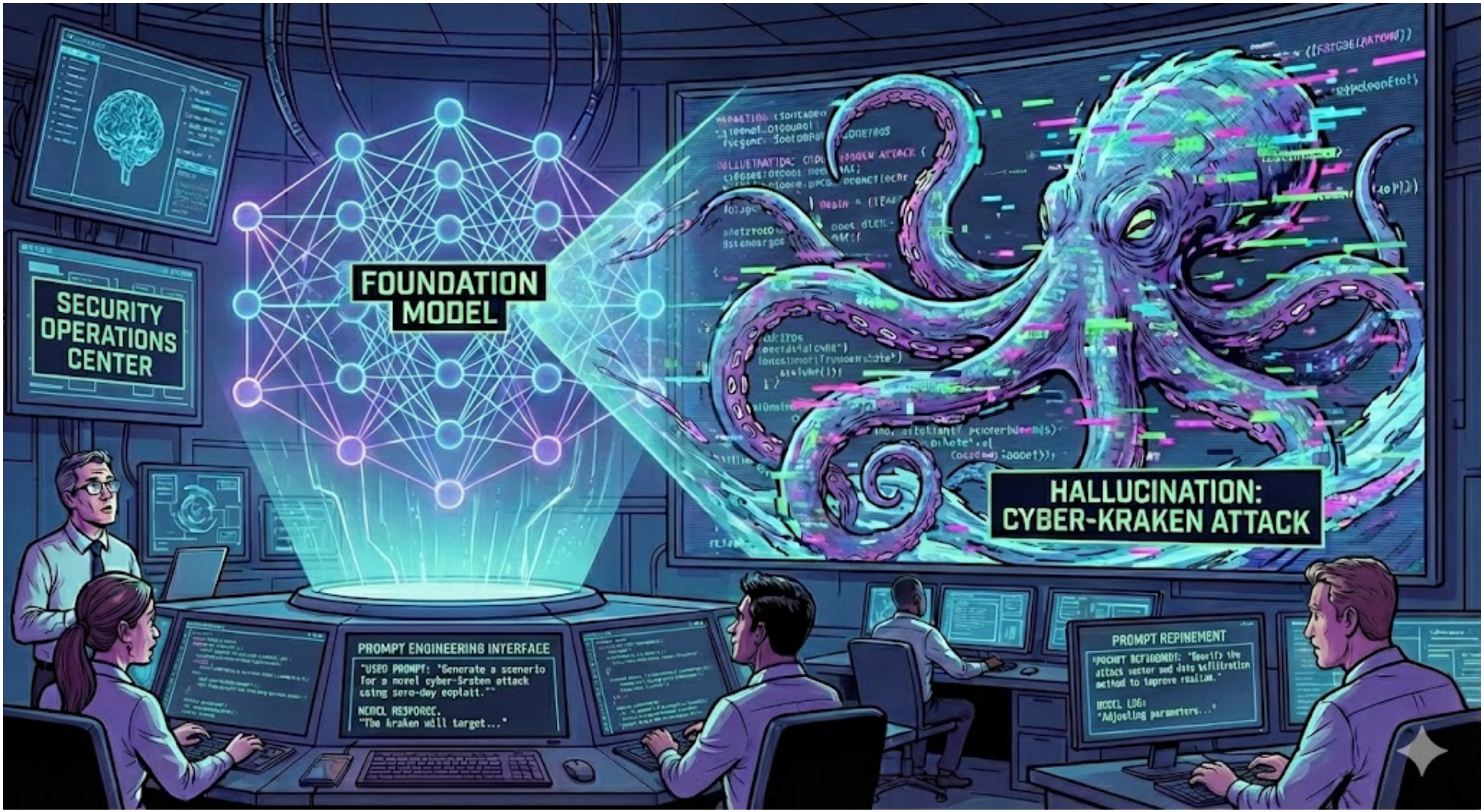
- Incident response is **managed by security experts**.
- We have a global shortage of more than 4 million experts.
- Pressing need for new decision support systems!

Next Generation Incident Response System

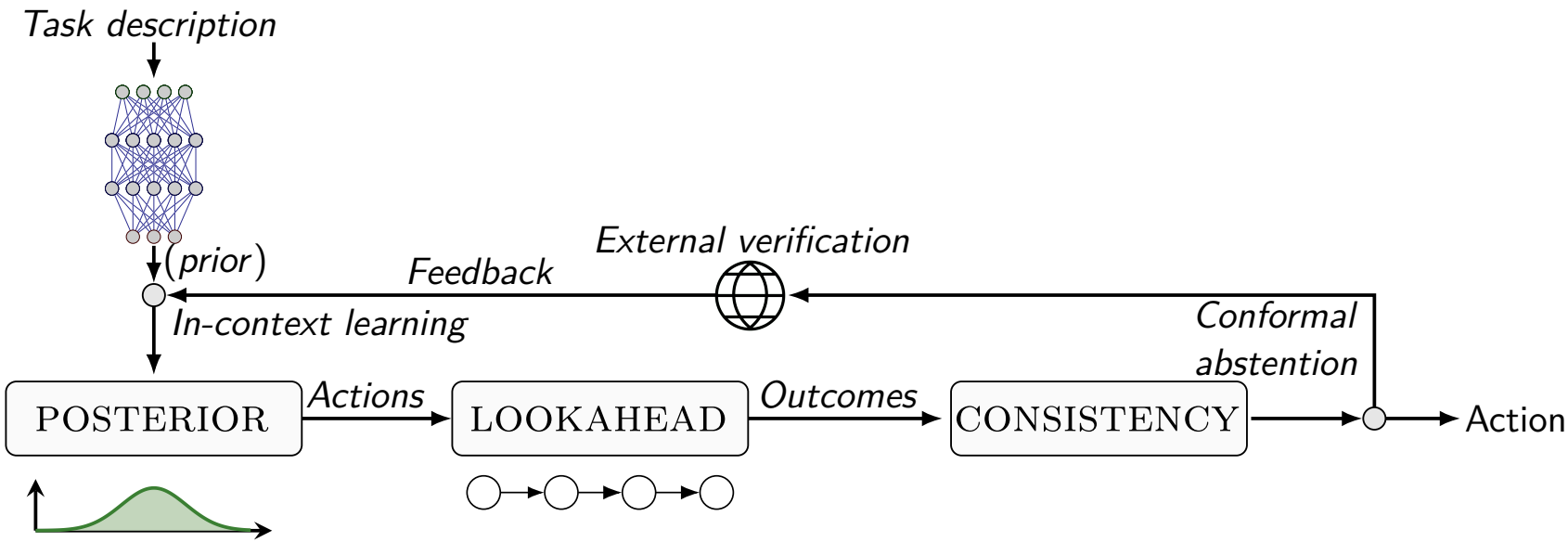


- We develop a response system centered around a lightweight foundation model.
- We analyze hallucination risks and establish theoretical reliability guarantees.

How to build a reliable system from unreliable components?

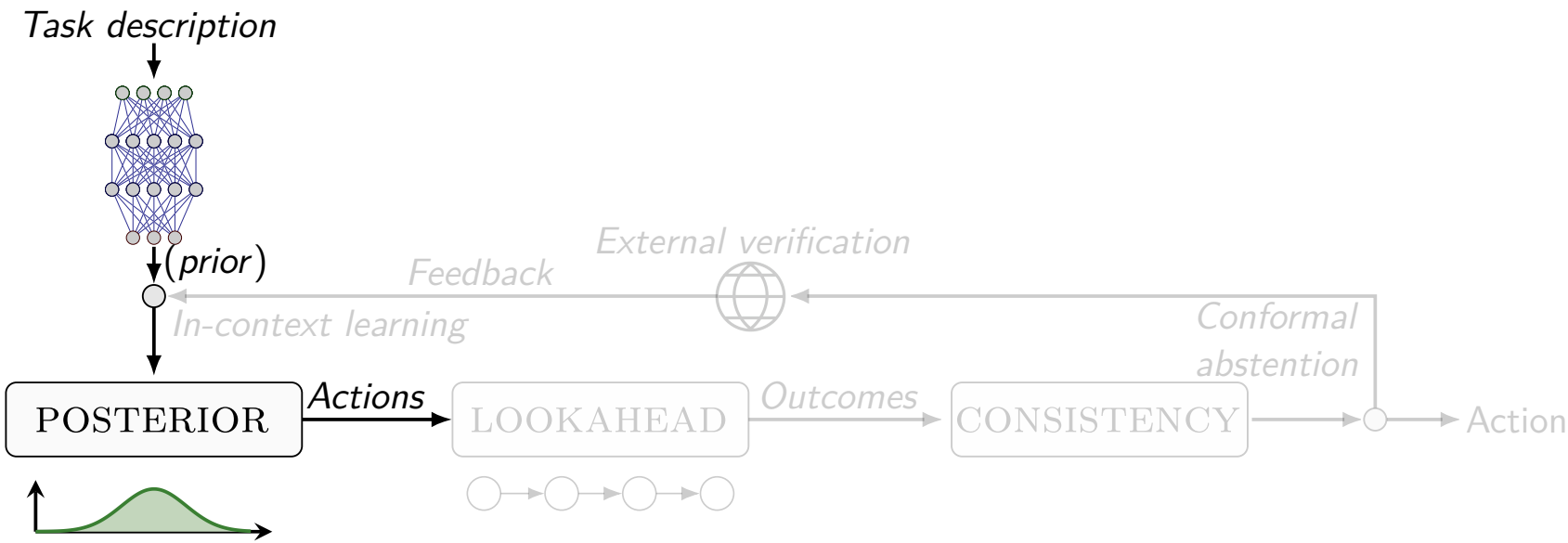


Incident Response Planning with a **Foundation Model**



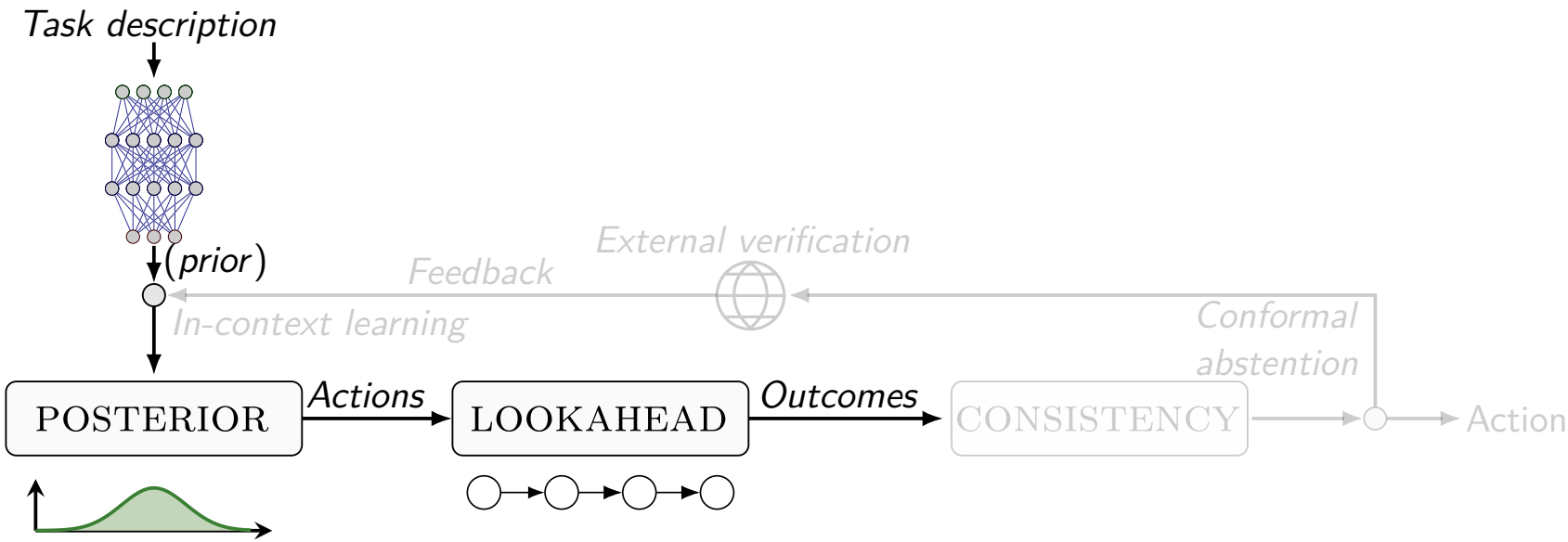
- We use the **model to generate candidate actions.**
- We evaluate actions through lookahead.
- We detect likely hallucinations by evaluating consistency.
- Abstain from actions with low consistency.
- Refine actions via **in-context learning** from feedback.

Incident Response Planning with a **Foundation Model**



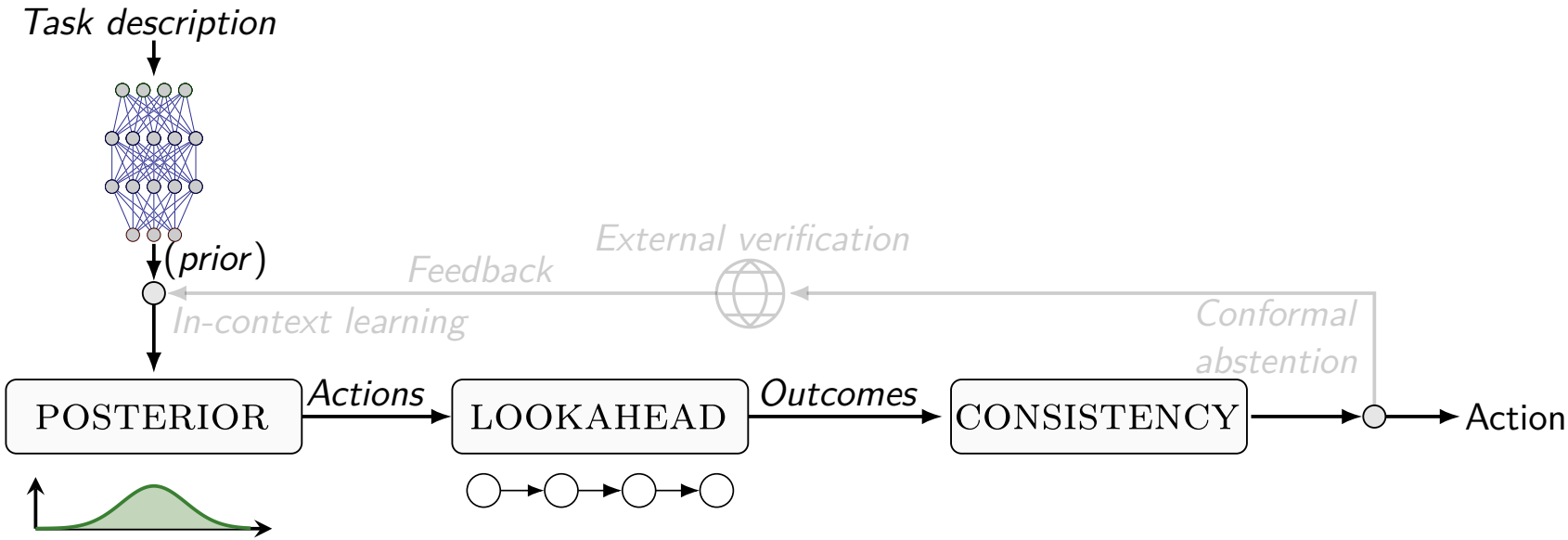
- We use the **model to generate candidate actions.**
- We evaluate actions through lookahead.
- We detect likely hallucinations by evaluating consistency.
- Abstain from actions with low consistency.
- Refine actions via **in-context learning** from feedback.

Incident Response Planning with a **Foundation Model**



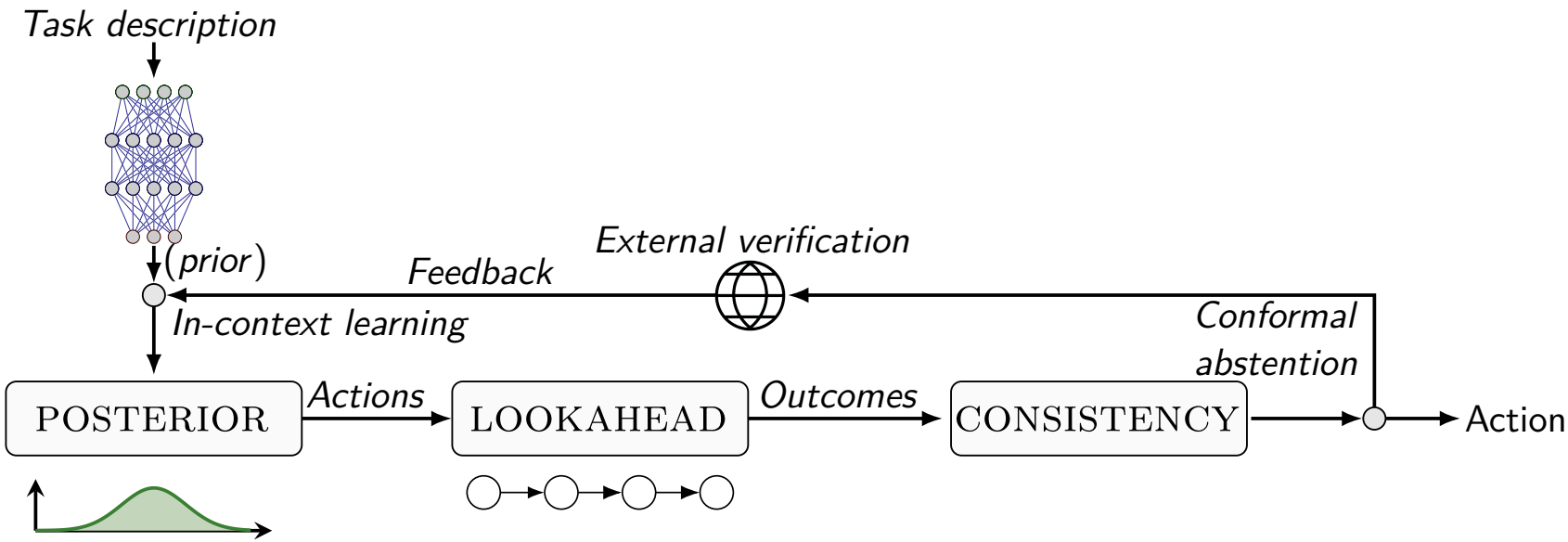
- We use the **model to generate candidate actions.**
- We evaluate actions through lookahead.
- We detect likely hallucinations by evaluating consistency.
- Abstain from actions with low consistency.
- Refine actions via **in-context learning** from feedback.

Incident Response Planning with a **Foundation Model**



- We use the **model to generate candidate actions.**
- We evaluate actions through lookahead.
- We detect likely hallucinations by evaluating consistency.
- Abstain from actions with low consistency.
- Refine actions via **in-context learning** from feedback.

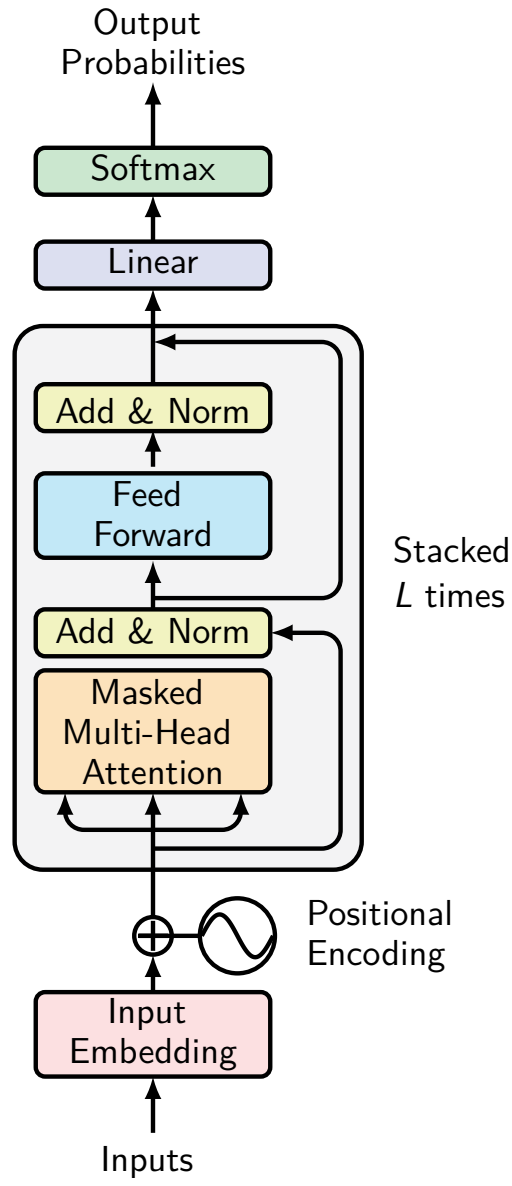
Incident Response Planning with a **Foundation Model**



- We use the **model to generate candidate actions.**
- We evaluate actions through lookahead.
- We detect likely hallucinations by evaluating consistency.
- Abstain from actions with low consistency.
- Refine actions via **in-context learning** from feedback.

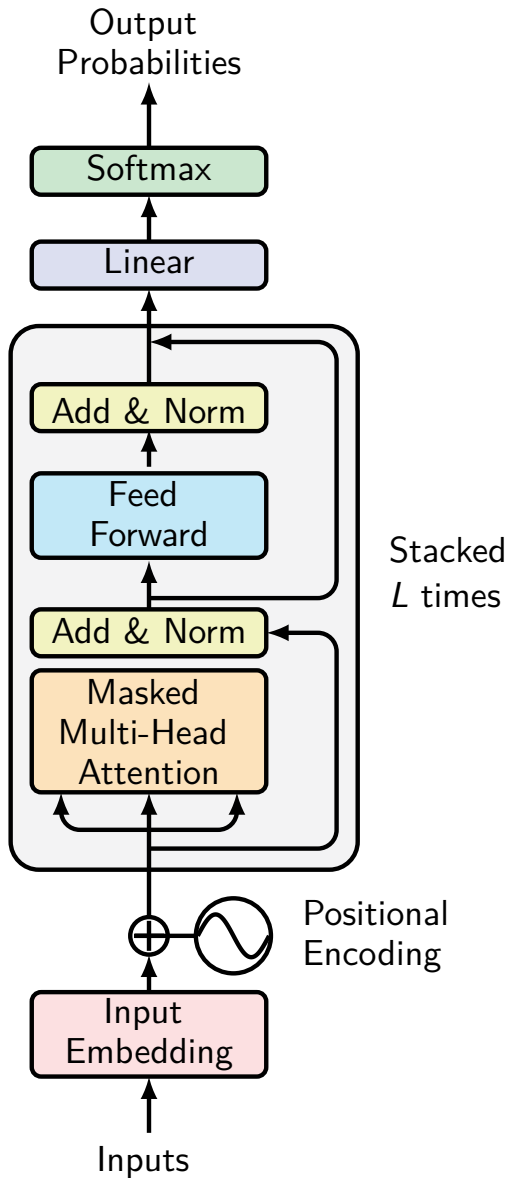
Different Types of **Foundation Models**

- Based on the transformer architecture.
- Trained on **vast datasets**.
- Billions of parameters.
- Examples:
  - Large language models (e.g., DeepSeek).
  - Time series models (e.g., Chronos).
  - Speech and audio models (e.g., Whisper).
  - Multi-modal models (e.g., Sora).



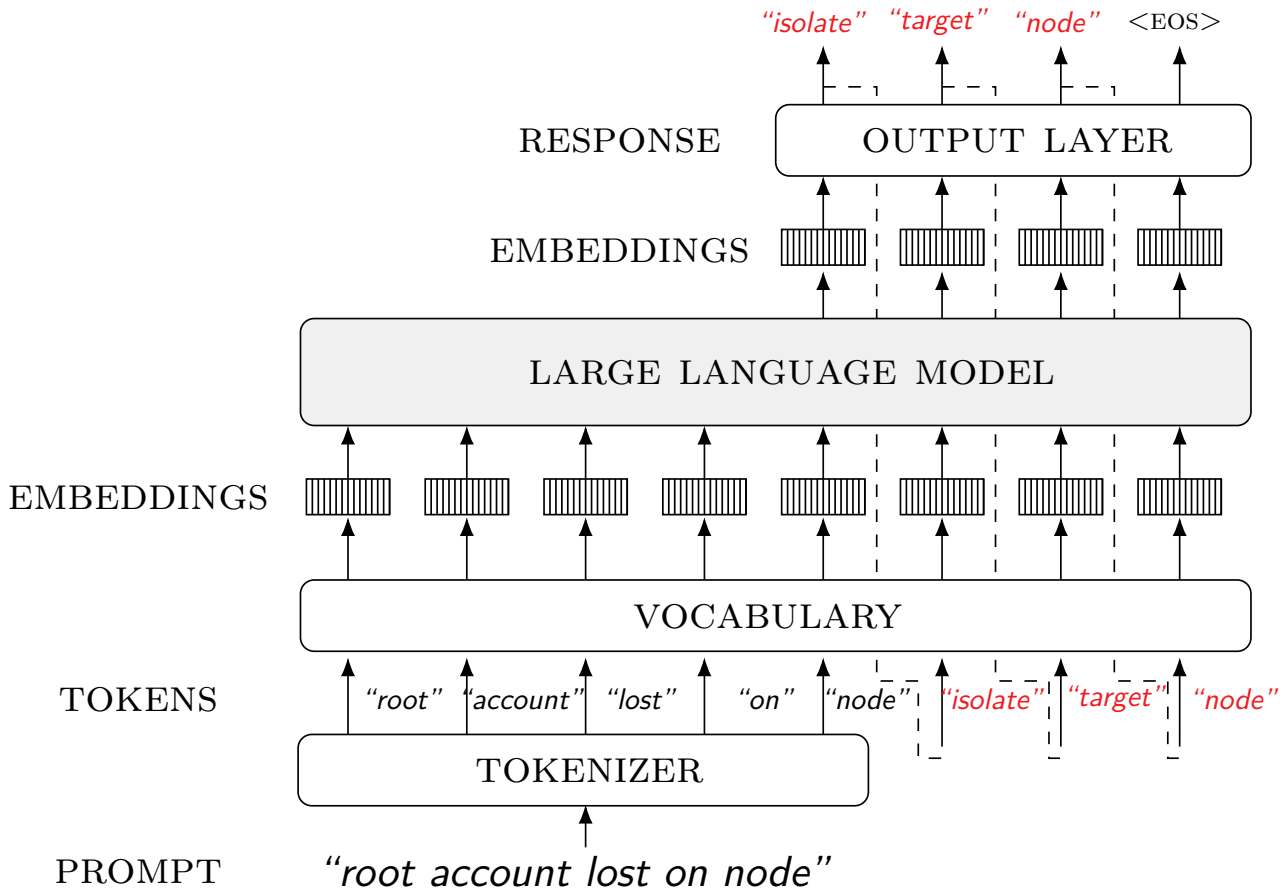
Different Types of **Foundation Models**

- Based on the transformer architecture.
- Trained on **vast datasets**.
- Billions of parameters.
- Examples:
  - Large language models (e.g., DeepSeek).
  - Time series models (e.g., Chronos).
  - Speech and audio models (e.g., Whisper).
  - Multi-modal models (e.g., Sora).

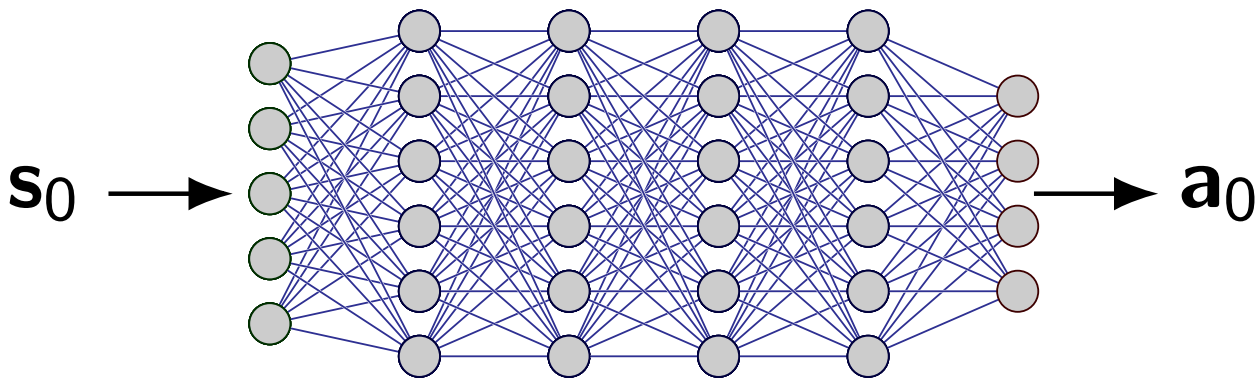


Generating Candidate Actions

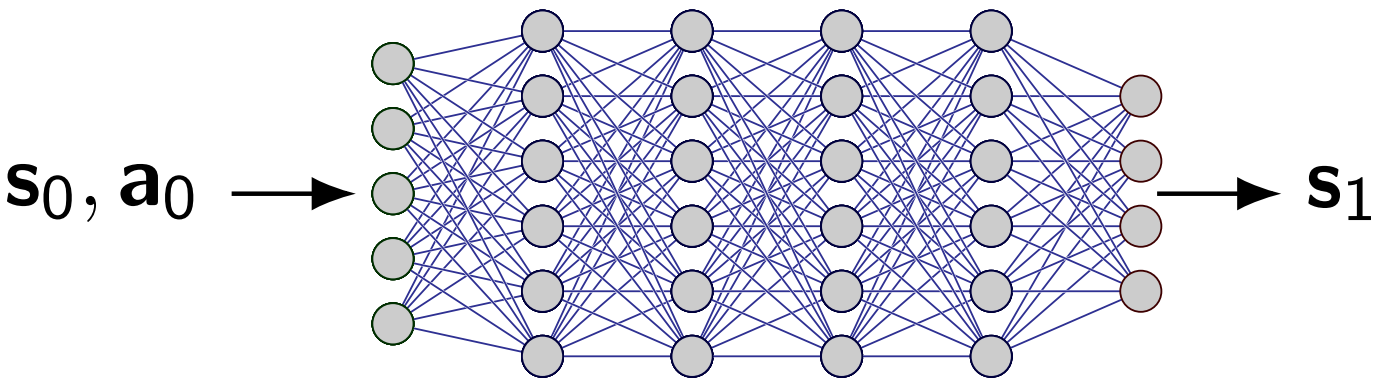
- Generate N candidate actions via **auto-regressive sampling.**
- Can think of the LLM as a base strategy.



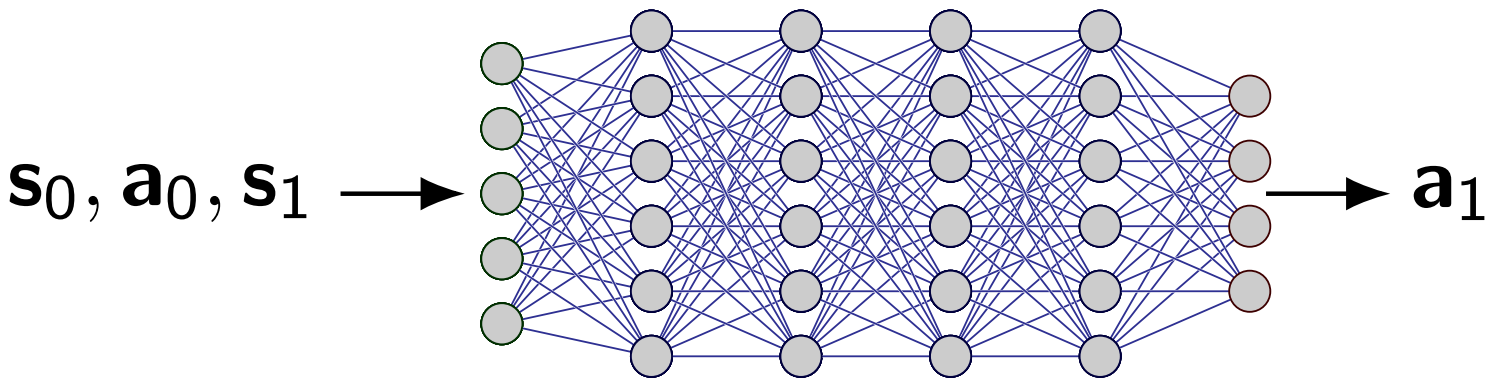
Lookahead Simulation with the LLM



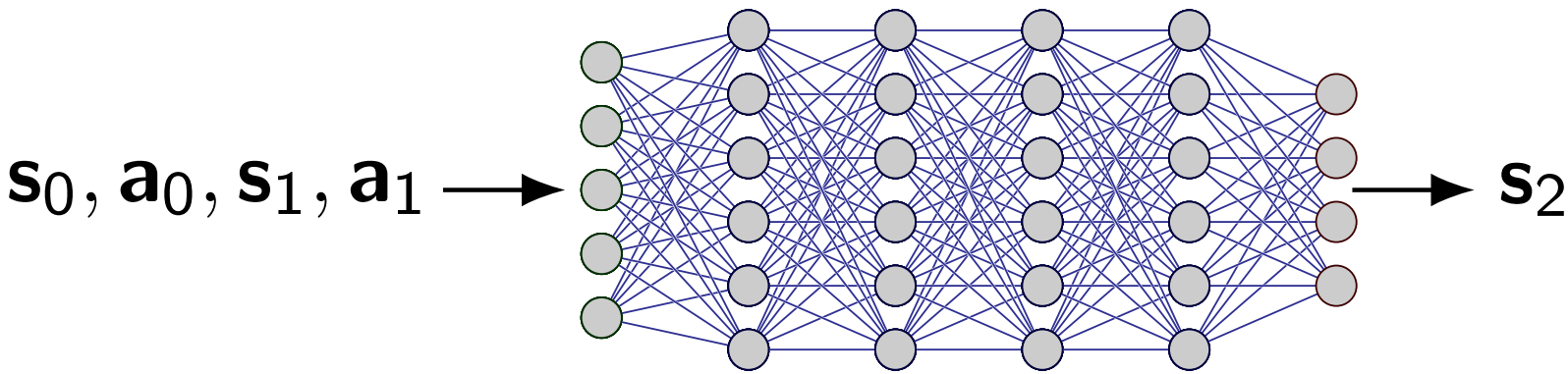
Lookahead Simulation with the LLM



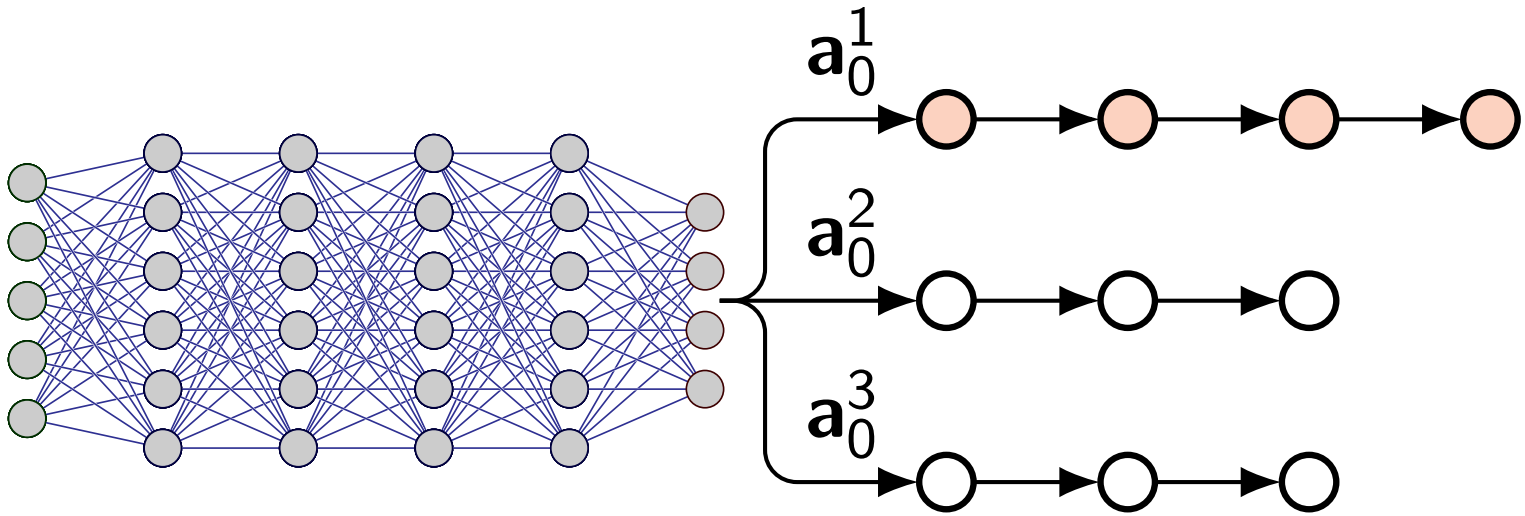
Lookahead Simulation with the LLM



Lookahead Simulation with the LLM



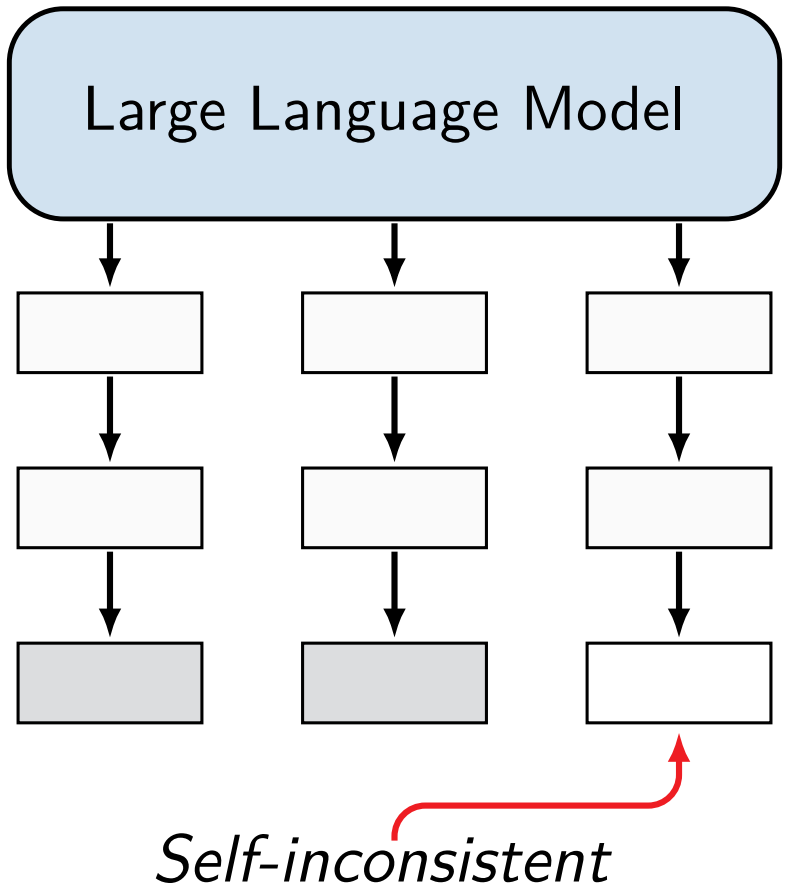
Lookahead Simulation with the LLM



- For each candidate action $\mathbf{a}_t^i$, we use the LLM to predict the subsequent states and actions.
- We select the action with the best outcome.

Evaluating the **Consistency** of Actions

- We use **inconsistency** as an indication of hallucination.



Abstaining from Inconsistent Actions

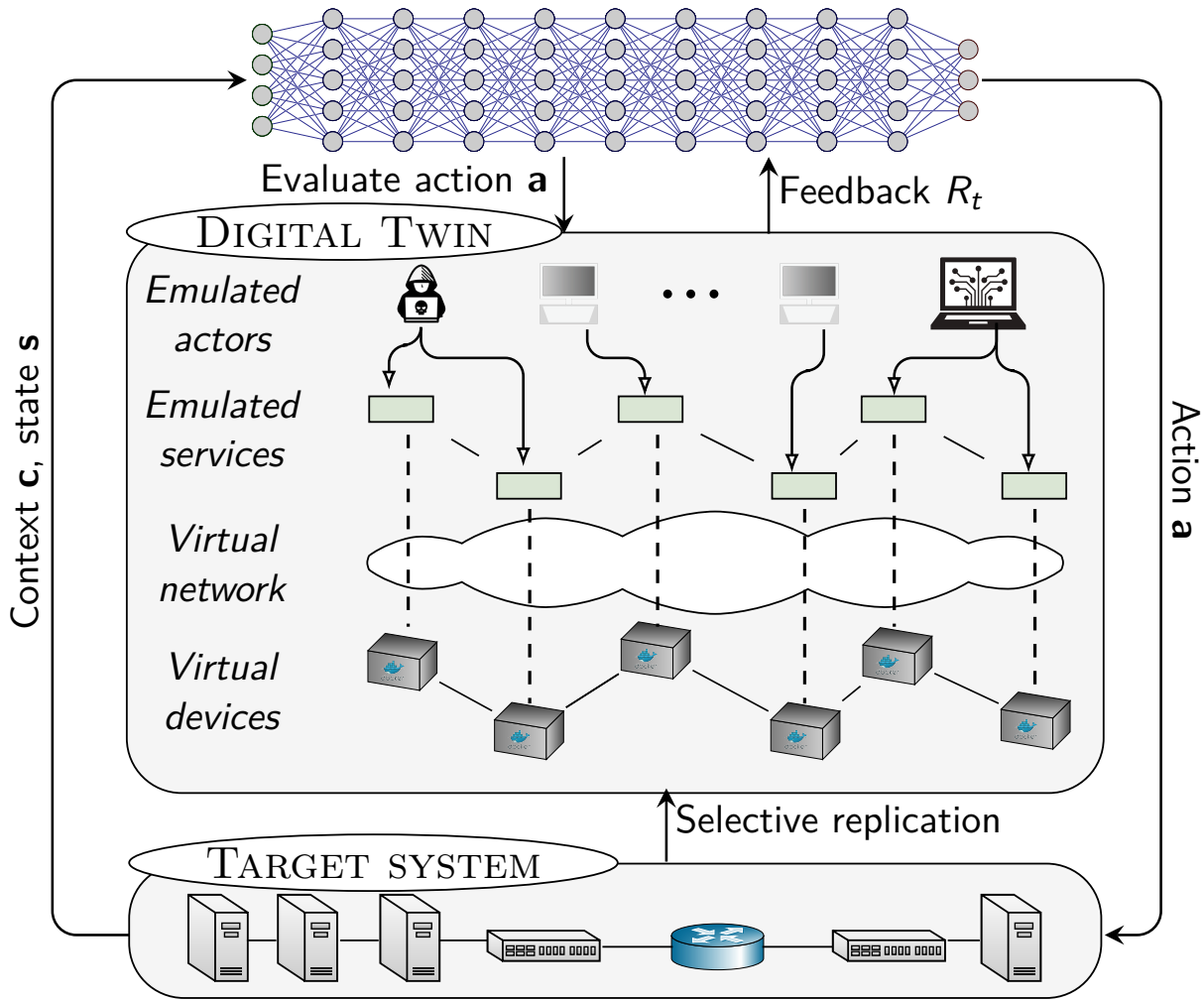
- Let $\lambda(\mathbf{a}) \in [0, 1]$ be a function that evaluates the consistency of a given action $\mathbf{a}$.
- We use this function to abstain from actions with low consistency, as expressed by the following decision rule:

$$\rho_\gamma(\mathbf{a}_t) = \begin{cases} 1 \text{ (abstain)}, & \text{if } \lambda(\mathbf{a}_t) \leq \gamma, \\ 0 \text{ (not abstain)}, & \text{if } \lambda(\mathbf{a}_t) > \gamma, \end{cases}$$

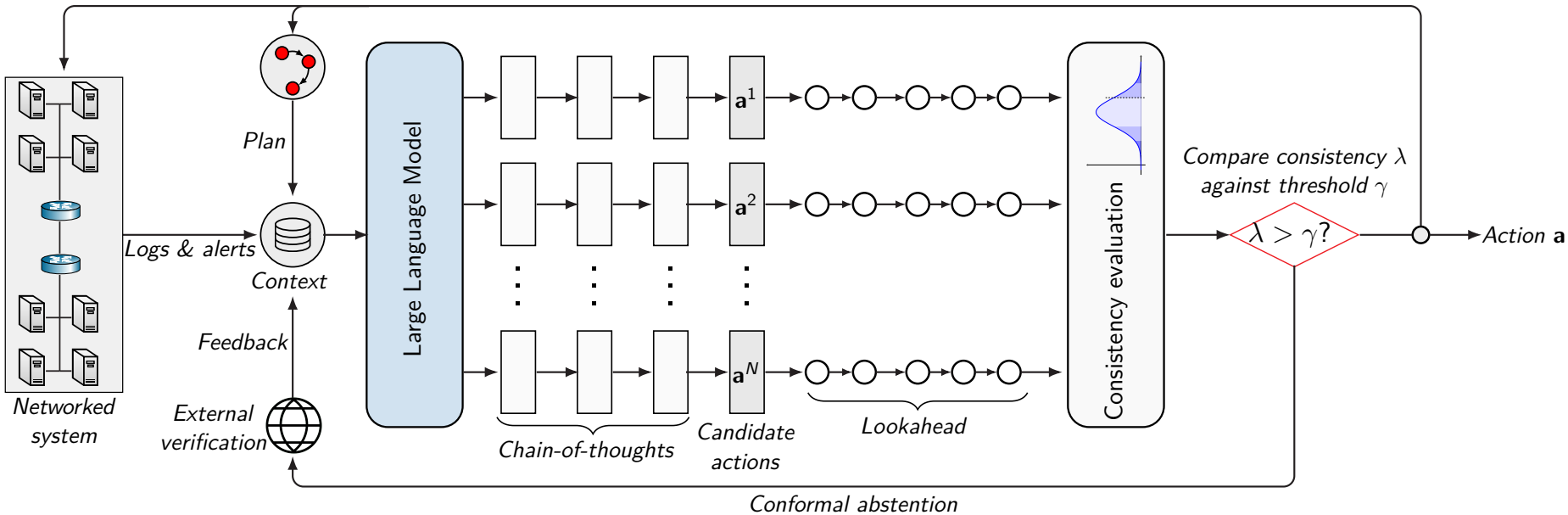
where $\gamma \in [0, 1]$ is a **consistency threshold**.

In-Context Learning from Feedback

If an action does not meet the **consistency threshold**, we abstain from it, collect external feedback (e.g., from a digital twin), and select a new action through in-context learning.



Summary of Our Framework



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.

Conformal Abstention

Let $\{\mathbf{a}_i\}_{i=1}^n$ be a *calibration dataset* of hallucinated actions.

Proposition 1 (Informal)

- *Assume the actions in the calibration dataset* $\{\mathbf{a}_i\}_{i=1}^n$ *are i.i.d.*
- *Let* $\tilde{\mathbf{a}}$ *be an hallucinated action from the same distribution.*
- *Let* $\kappa \in (0, 1]$ *be a desirable upper bound on the hallucination probability.*

Define the threshold

$$\tilde{\gamma} = \inf \left\{ \gamma \;\middle|\; \frac{|\{i \mid \lambda(\mathbf{a}_i) \leq \gamma\}|}{n} \geq \frac{\lceil (n+1)(1-\kappa) \rceil}{n} \right\},$$

where $\lceil \cdot \rceil$ *is the ceiling function. We have*

$$P\left(\textit{not abstain from } \tilde{\mathbf{a}}\right) \leq \kappa.$$

Regret Bound for In-Context Learning

Proposition 2 (Informal)

- *Let $\mathcal{R}_K$ denote the **Bayesian regret.***
- *Assume that the LLM's output distribution is aligned with the posterior given the context.*
- *Assume bandit feedback.*

We have

$$\mathcal{R}_K \leq C\sqrt{|\mathcal{A}| K \ln K},$$

where $C > 0$ is a universal constant, $\mathcal{A}$ is the set of actions, and K is the number of ICL iterations.

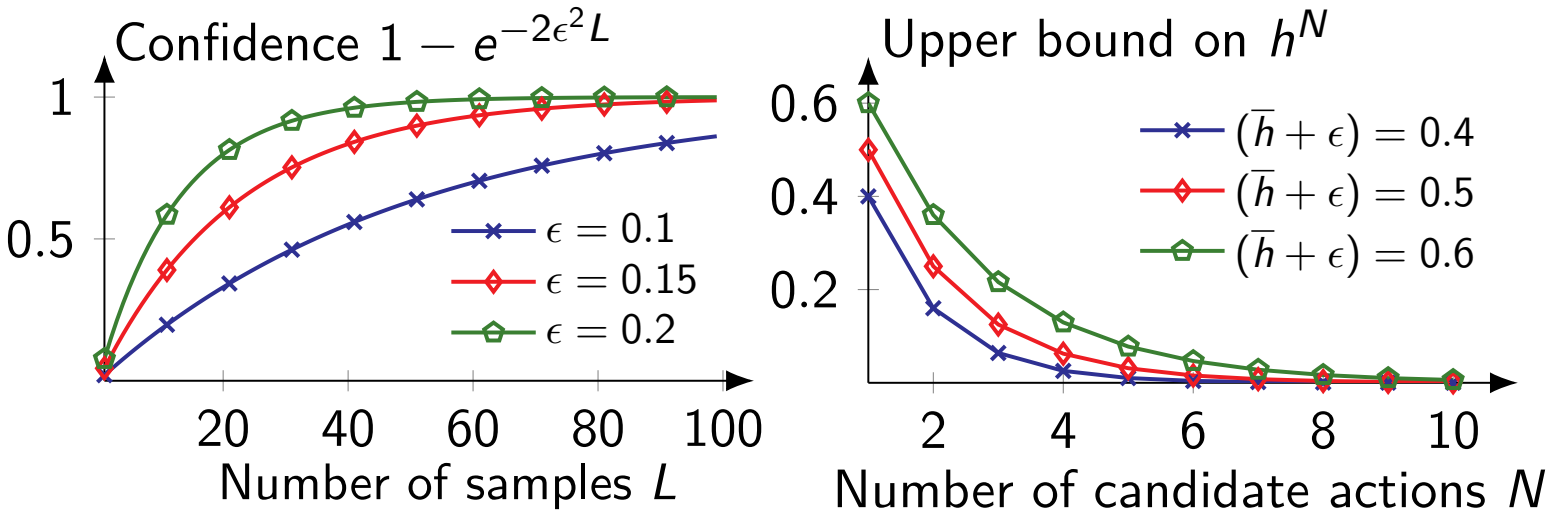
Chernoff Bound on the Hallucination Probability

Proposition 3 (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 4 (Informal)

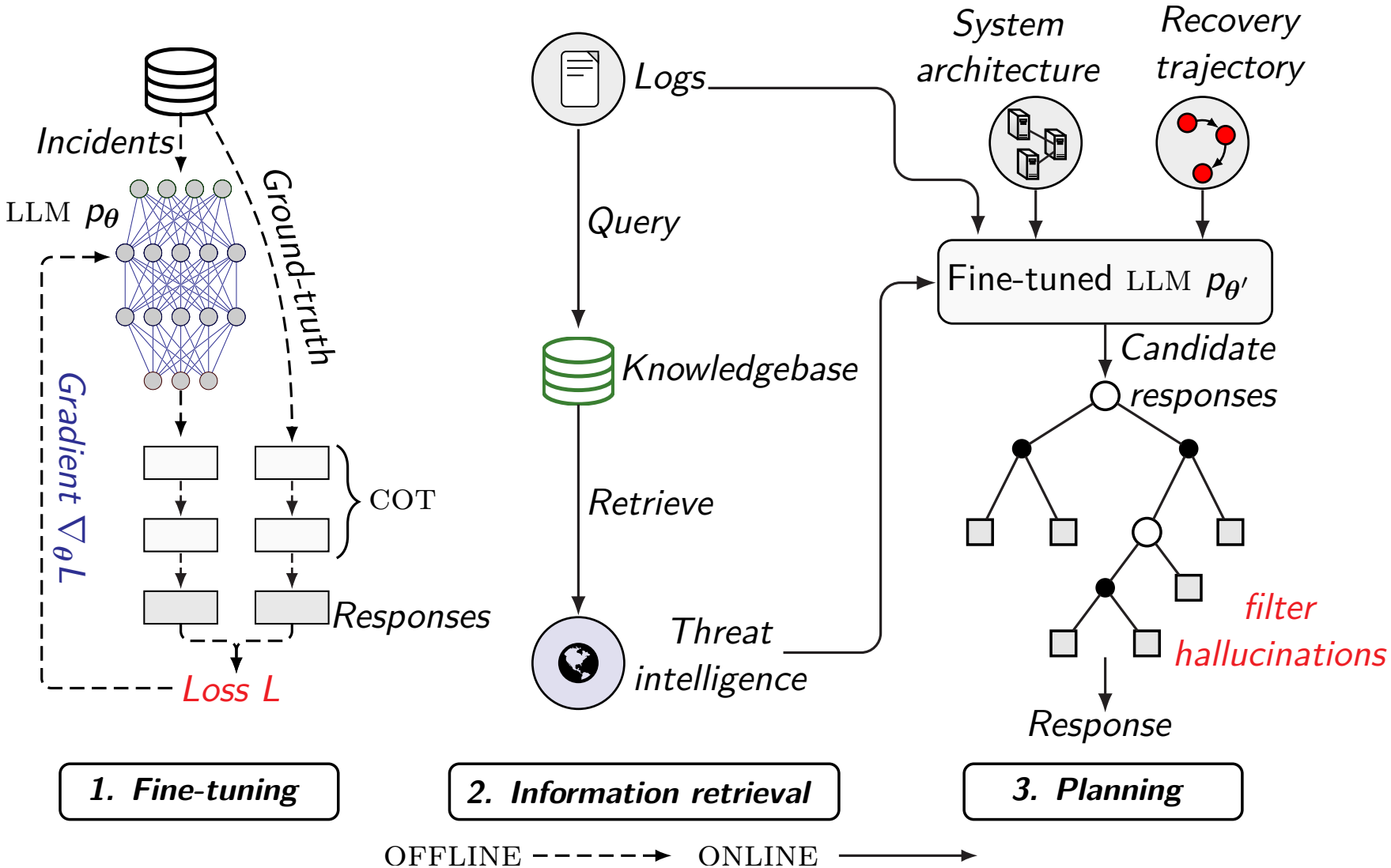
- *Let η be the total variation between LLM's predictions and true system dynamics.*
- *Let δ be the minimal **difference in recovery time between a hallucinated and non-hallucinated action.***
- *Assume at least one candidate action is not hallucinated.*

If

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

then the selected action will not be hallucinated.

Experiment Setup

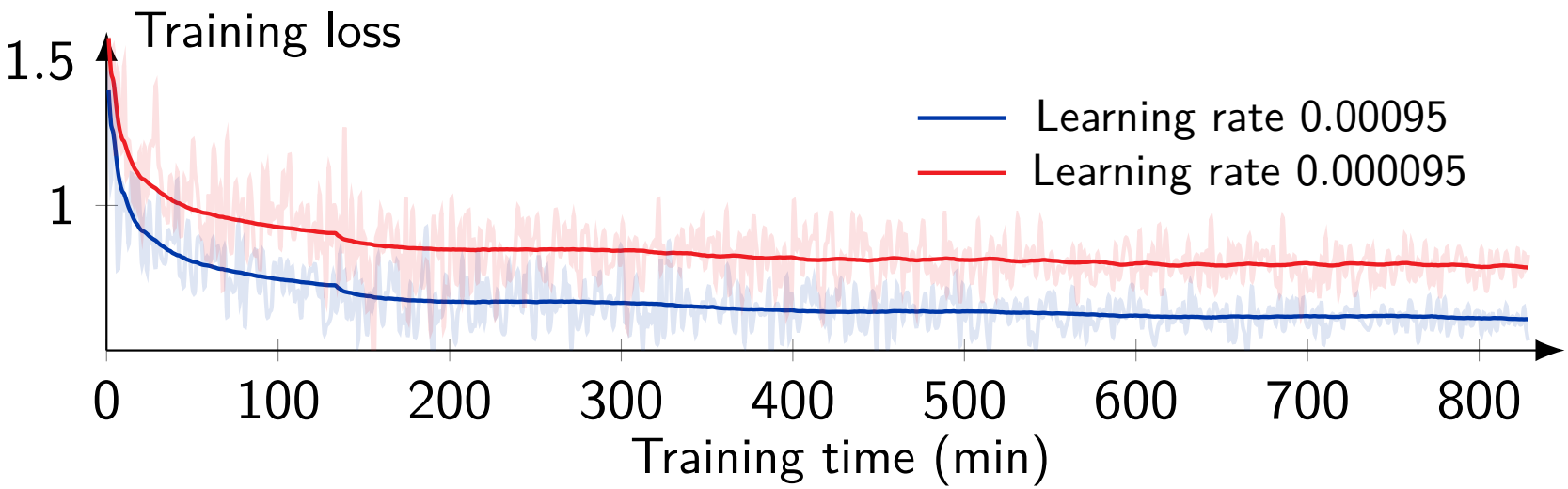


Instruction Fine-Tuning

- 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:**

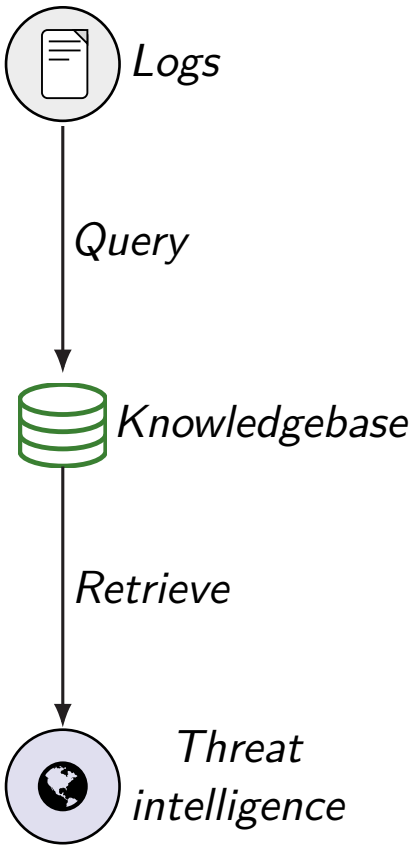
$$L = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{m_i} \ln p_\theta\left(\mathbf{y}_k^i \mid \mathbf{x}^i, \mathbf{y}_1^i, \ldots, \mathbf{y}_{k-1}^i\right),$$

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



Retrieval-Augmented Generation (RAG)

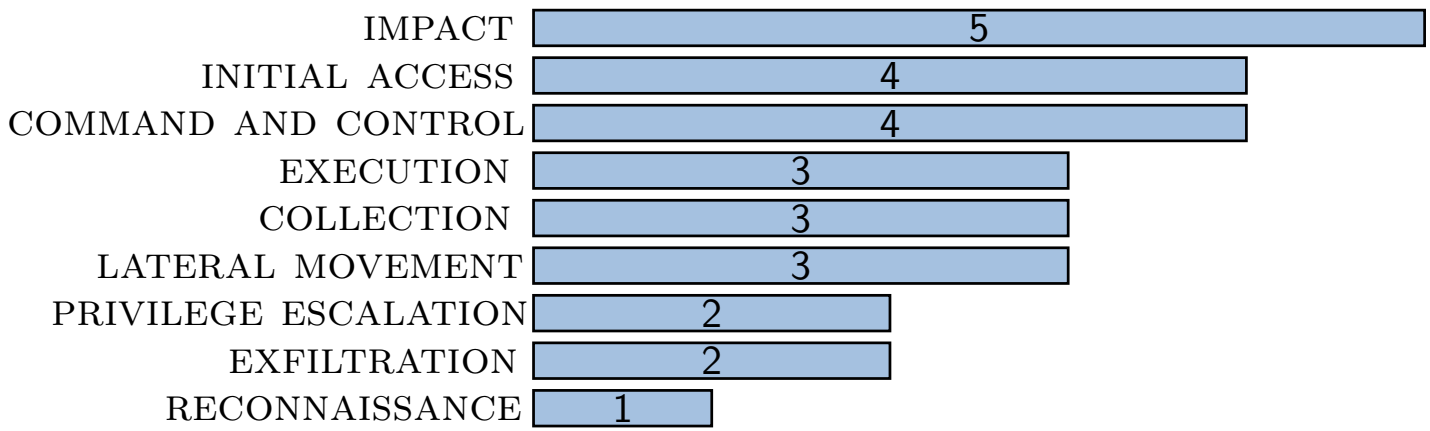
- We use regular expressions to extract indicators of compromise (IOC) from logs.
  - e.g., IP addresses, vulnerability identifiers, etc.
- We use the IOCs to **retrieve information about the incident** from public threat intelligence APIs, e.g., OTX.
- We include the retrieved information in the context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



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.



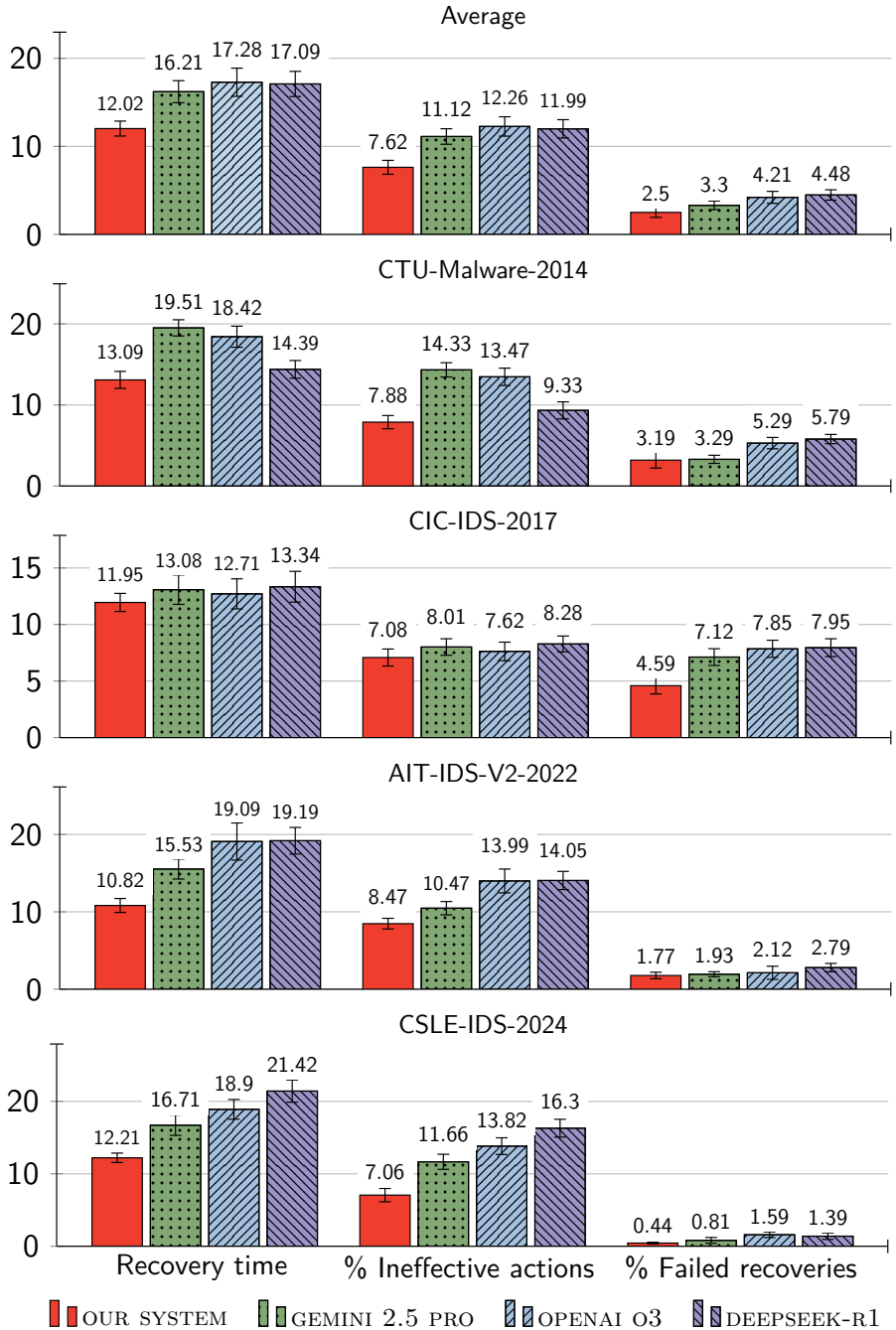
Distribution of MITRE ATT&CK tactics in the evaluation datasets.

Baselines

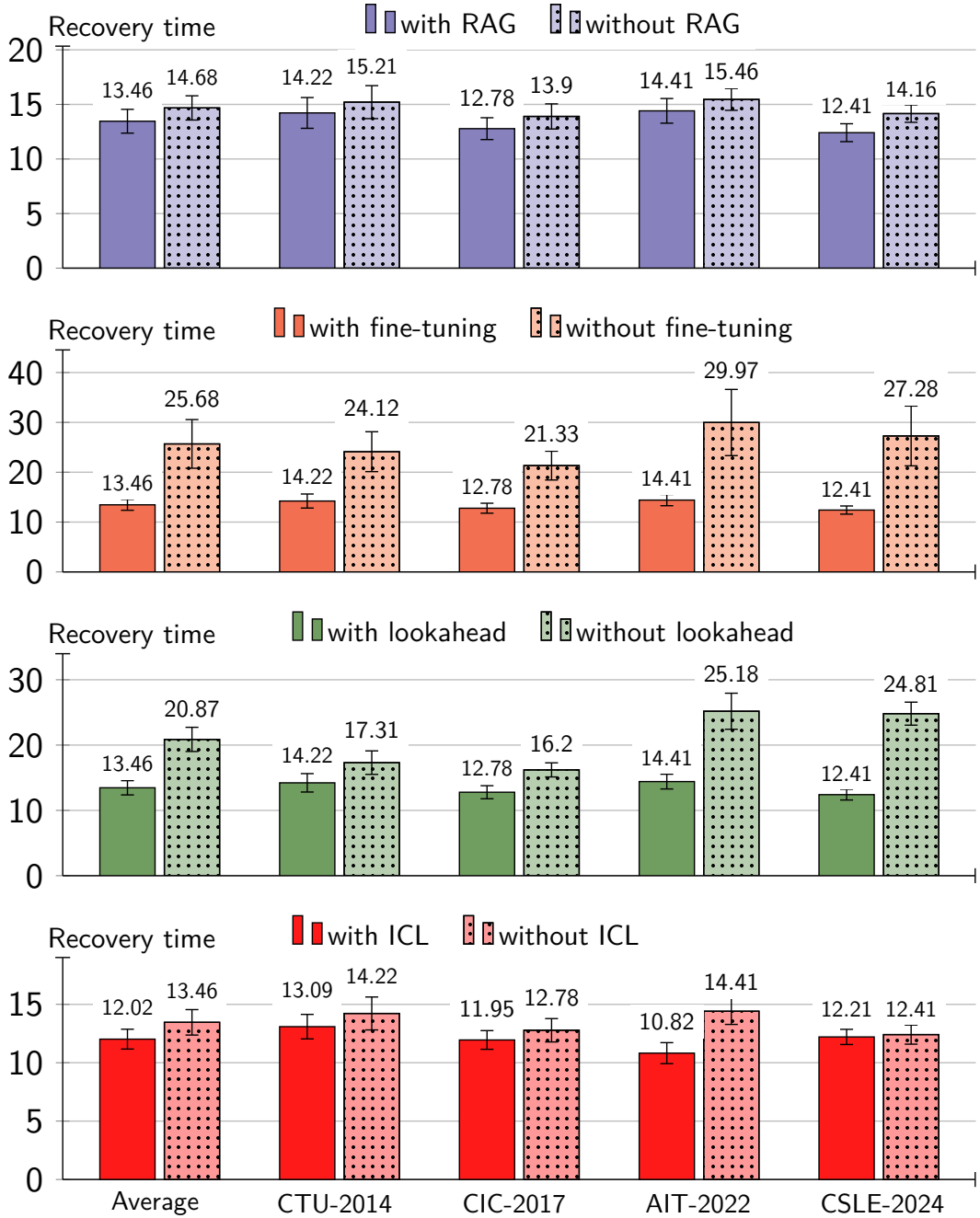
- We compare our system against **frontier LLMs.**
- Compared to the frontier models, our system is lightweight.

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

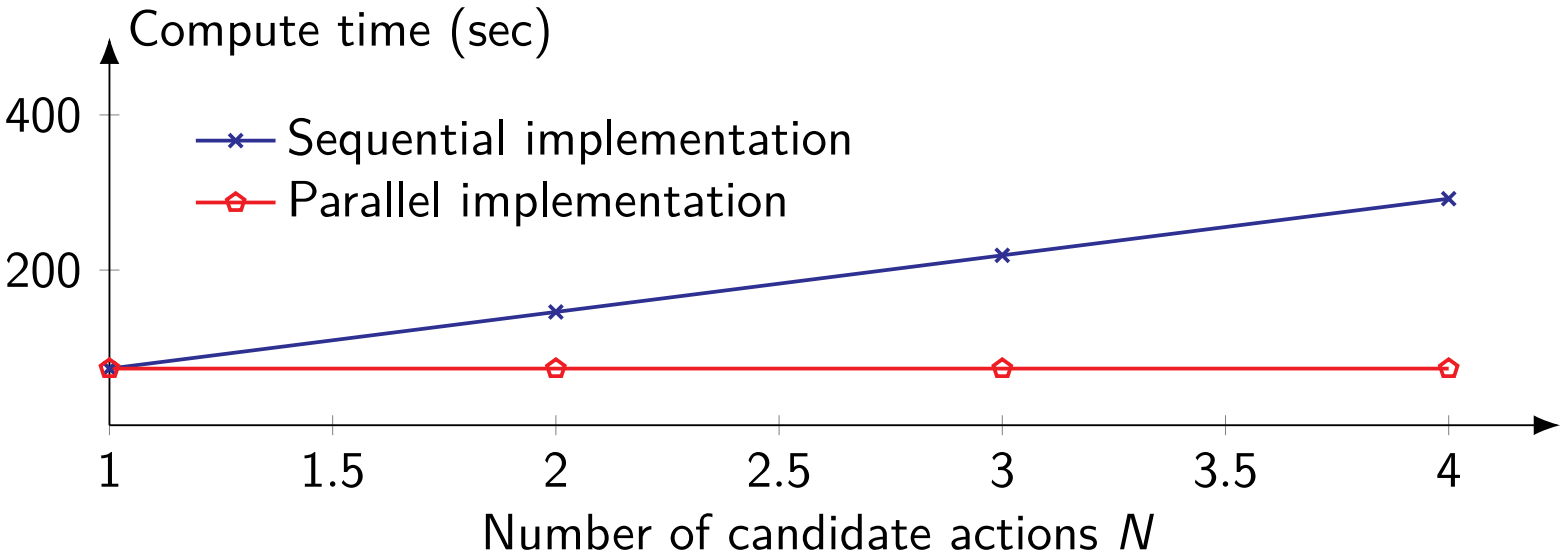
Evaluation Results



Ablation Study



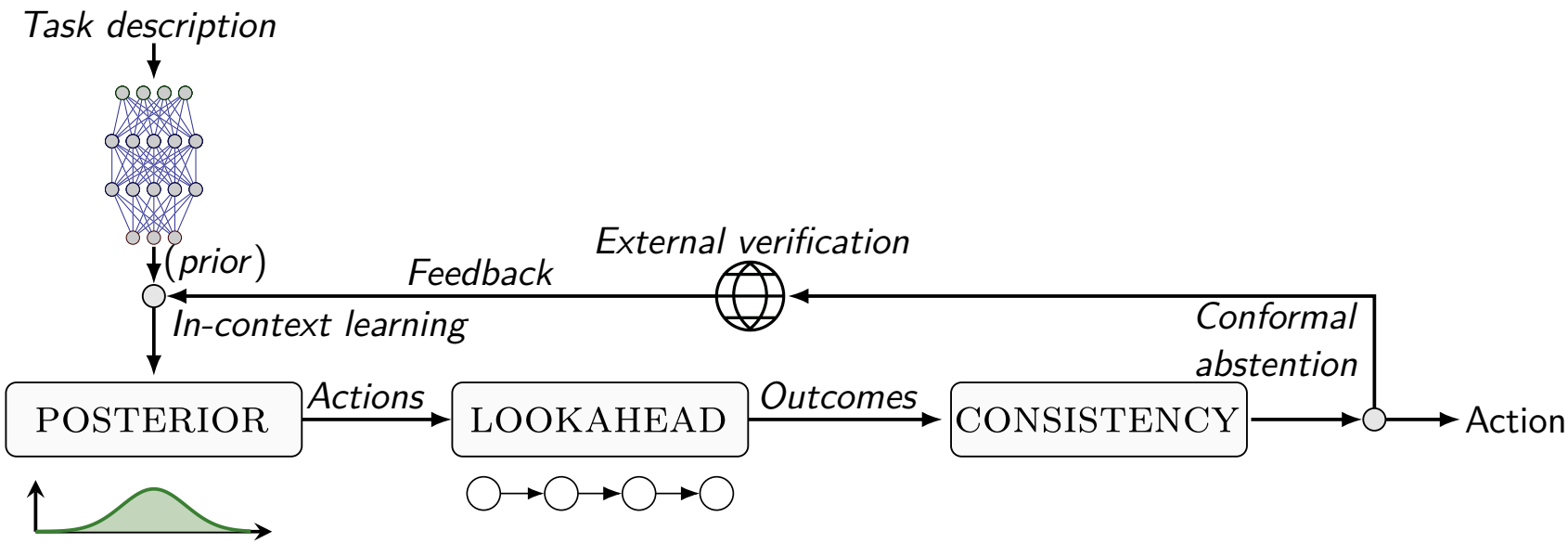
Scalability



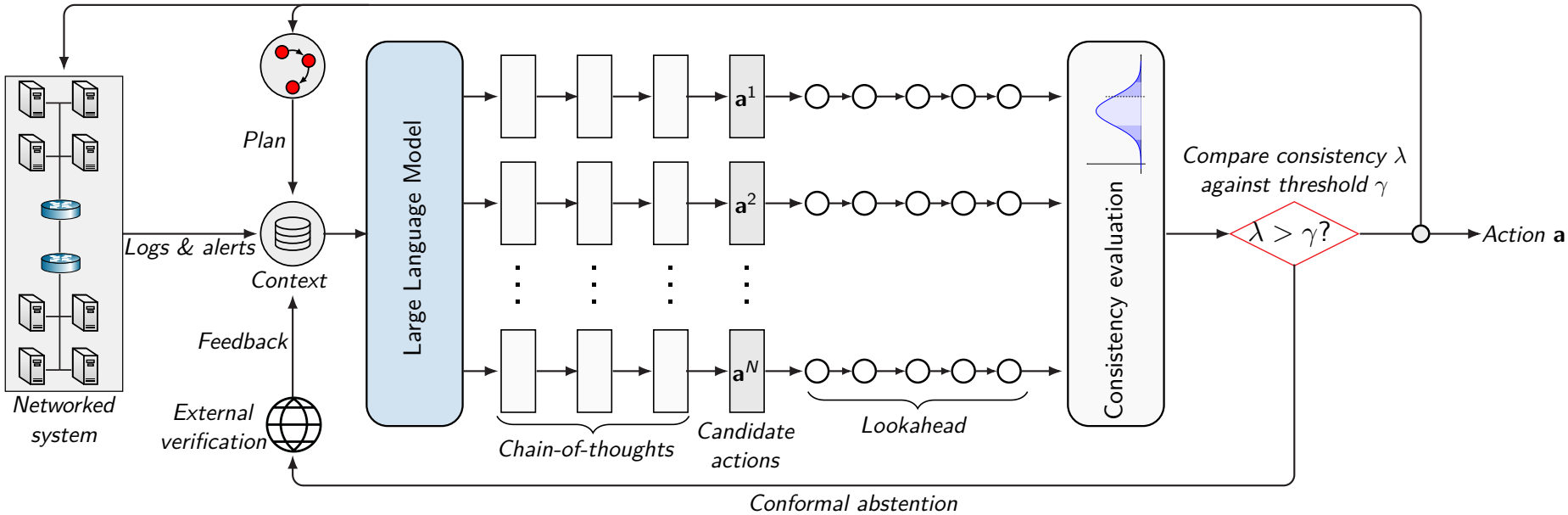
- The lookahead optimization is computationally intensive since it requires making multiple inferences with the LLM.
- The computation can be parallelized across multiple GPU.

Conclusion

- **Foundation models will play a key role in cybersecurity.**
  - Effective at tackling the scalability challenge.
  - Remarkable knowledge management capabilities.
- We present a framework for security planning.
  - Allows to control the hallucination probability.
  - Significantly outperforms frontier LLMs.



References



- Video demonstration:
  - https://www.youtube.com/watch?v=SCxq2ye-R4Y
- Code:
  - https://github.com/Kim-Hammar/llm_incident_response_ndss26
- Dataset and model weights:
  - https://huggingface.co/datasets/kimhammar/CSLE-IncidentResponse-V1
  - https://huggingface.co/kimhammar/LLMIncidentResponse