LLM for Network Management

Xiaohui Xie
Tsinghua University
xiexiaohui@tsinghua.edu.cn

Background & Motivation

- **Challenges:**
  - Complex environments
  - Diverse demands
  - Rapid iterations
- **LLMs' Emerging Capabilities:**
  - Logical Reasoning
  - Generalization
  - Tool Utilization

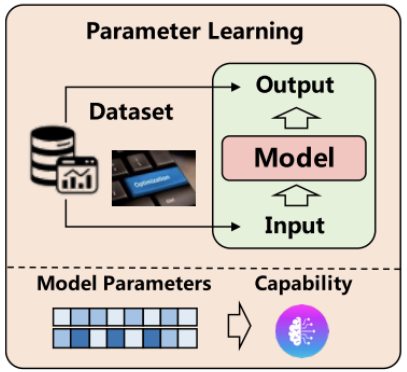
Based on Assumptions

$$QoE = \alpha*\log(R) - \beta*\log(D) - \gamma*\log(C)$$

R: Resolution
D: Delay
C: Change rate

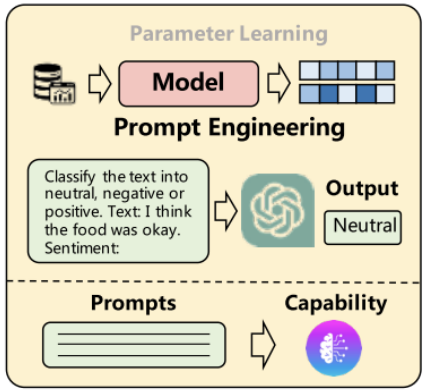
Mathematical Modeling

Domain dataset

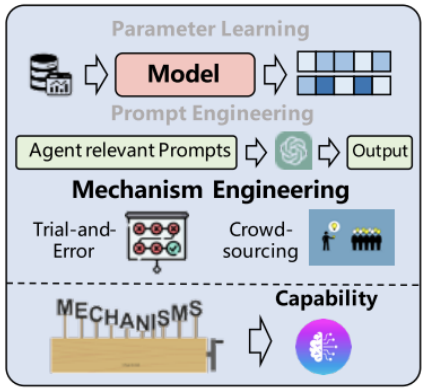


Machine Learning

Extensive datasets and Large-scale parameters



LLM



LLM Agent

Increasing Attention

- **HotNets 2023**
  - November 2023 @ MIT
  - Two sessions (2 out of 9) are dedicated to this topic.

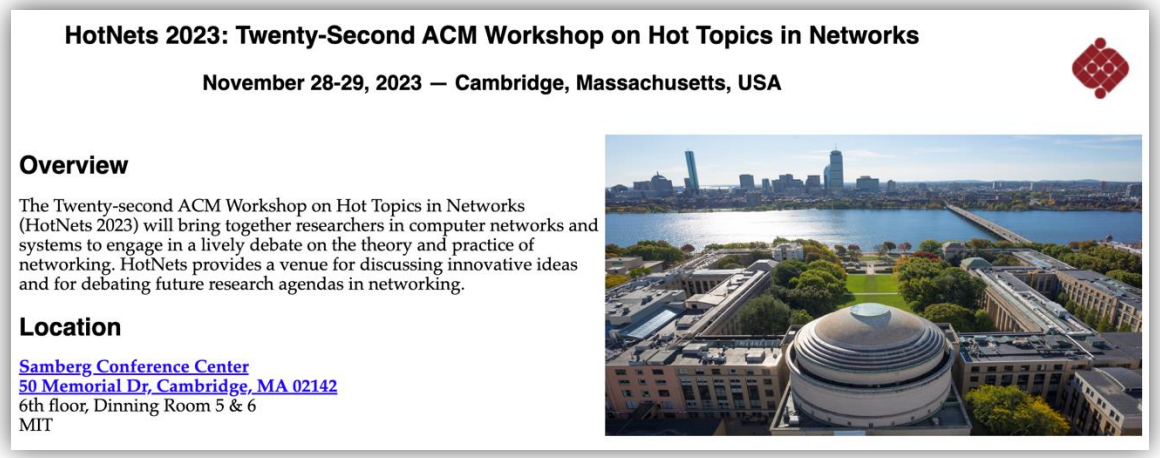
Session 6: Can LLMs Manage Networks?
Session Chair: Nate Foster (Cornell)

Adapting Foundation Models for Operator Data Analytics
Manikanta Kotaru (Microsoft)

A Holistic View of AI-driven Network Incident Management
Pouya Hamadanian (Microsoft Research, MIT); Behnaz Arzani, Sadjad Fouladi, Siva Kesava
Rodrigo Fonseca (Azure Systems Research); Denizcan Billor, Ahmad Cheema, Edet Nkposo
(Microsoft Research)

What do LLMs need to Synthesize Correct Router Configurations?
Rajdeep Mondal, Alan Tang (UCLA); Ryan Beckett (Microsoft Research); Todd Millstein, Ge

Enhancing Network Management Using Code Generated by Large Language Models
Sathiya Kumaran Mani (Microsoft); Yajie Zhou (Microsoft and Boston University); Kevin Hs
Segarra (Microsoft and Rice University); Trevor Eberl, Eliran Azulai, Ido Frizler, Ranveer Ch



Session 2: Can LLMs reason about networking problems, and their solution?
Session Chair: Ranjita Bhagwan (Google)

Towards Interactive Research Agents for Internet Incident Investigation
Yajie Zhou, Nengneng Yu (Boston University); Zaoxing Liu (University of Marylanc

PROSPER: Extracting Protocol Specifications Using Large Language Models
Prakhar Sharma, Vinod Yegneswaran (SRI International)

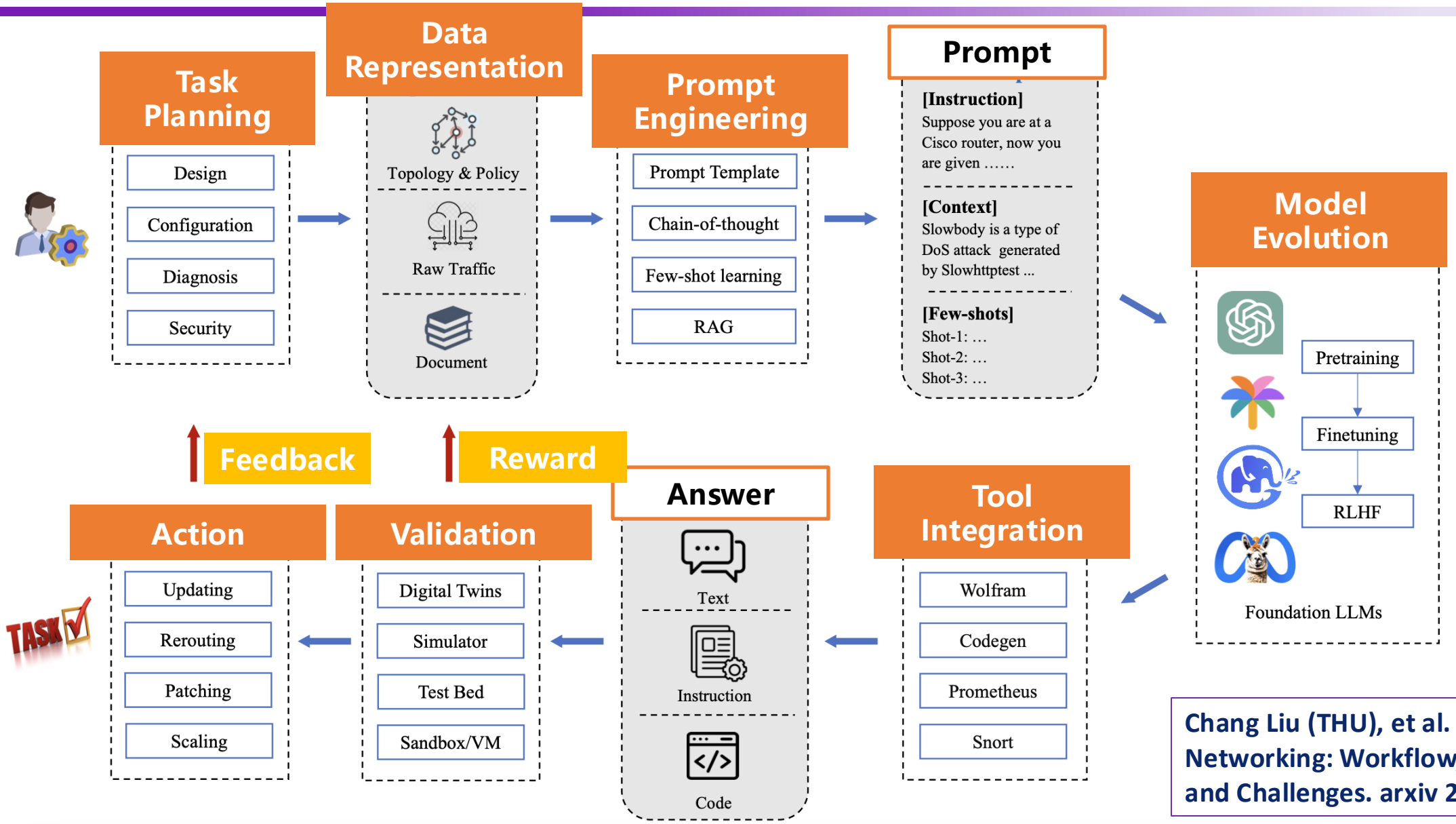
Towards Integrating Formal Methods into ML-Based Systems for Networking
Fengchen Gong, Divya Raghunathan, Aarti Gupta, Maria Apostolaki (Princeton Un

Toward Reproducing Network Research Results Using Large Language Models
Qiao Xiang, Yuling Lin, Mingjun Fan, Bang Huang, Siyong Huang, Ridi Wen (Xiam
Kong (Shanghai Jiao Tong University, China); Jiwu Shu (Xiamen University)

Increasing Attention

Institution	Research Paper	Conference
CUHK-Shenzhen	NetLLM: Adapting Large Language Models for Networking	SIGCOMM 24
ByteDance	NetAssistant: Dialogue Based Network Diagnosis in Data Center Networks	NSDI 24
NUS	Large Language Model guided Protocol Fuzzing	NDSS 24
BUPT	Following the Compass: LLM-Empowered Intent Translation with Manual Guidance	ICNP 24
Northeastern University	ConfigTrans: Network Configuration Translation Based on Large Language Models and Constraint Solving	ICNP 24
Huawei	NetConfEval: Can LLMs Facilitate Network Configuration?	CoNEXT 24
Microsoft & UIUC	Automatic Root Cause Analysis via Large Language Models for Cloud Incidents.	EuroSys 24

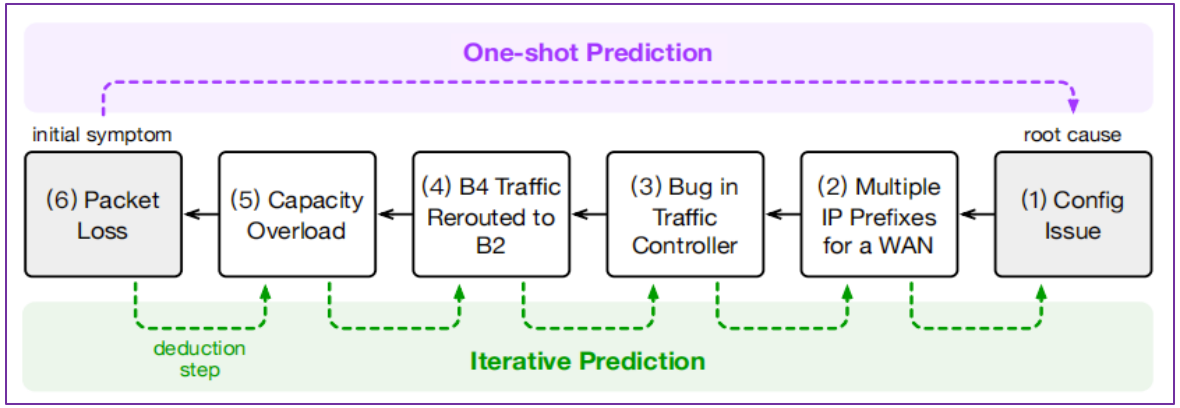
Workflow



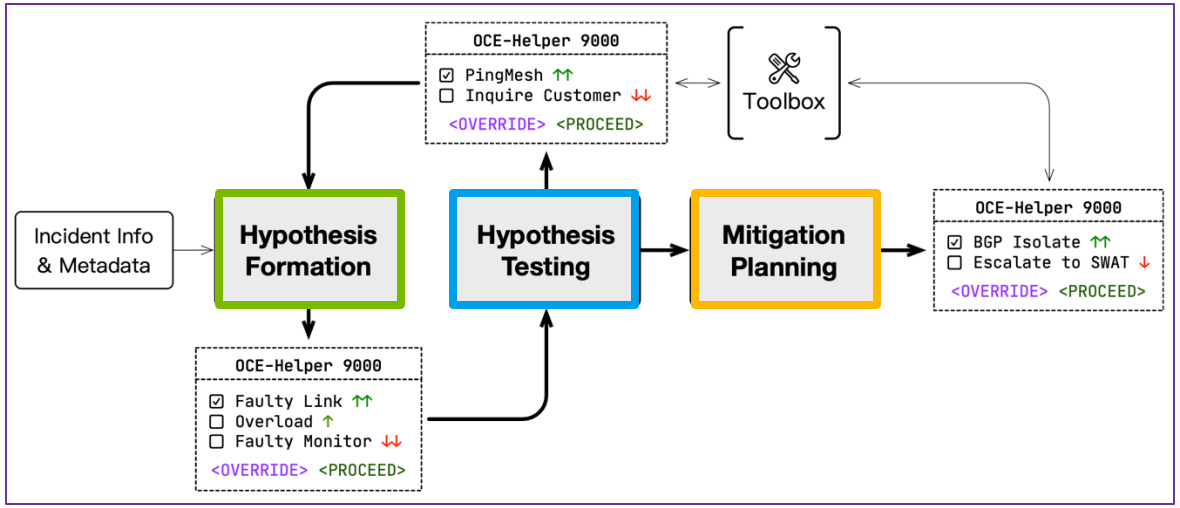
Chang Liu (THU), et al. LLM for Networking: Workflow, Advances and Challenges. arxiv 2024

LLM Agent as On-Call Engineer

- **Problems and Challenges**
  - Network failures occur frequently, and the cost of manual handling is high.
  - Single LLM operate as black boxes, making it difficult to accurately perform root cause analysis and troubleshooting of network failures.
- **Approach**
  - Decompose complex tasks, shifting from one-shot prediction to **multi-step iterative** predictions.
  - One round iteration = **hypothesis formation** + **hypothesis testing** + **mitigation planning**
  - Each stage is executed by dedicated agents, optimizing the next round of iteration based on execution feedback



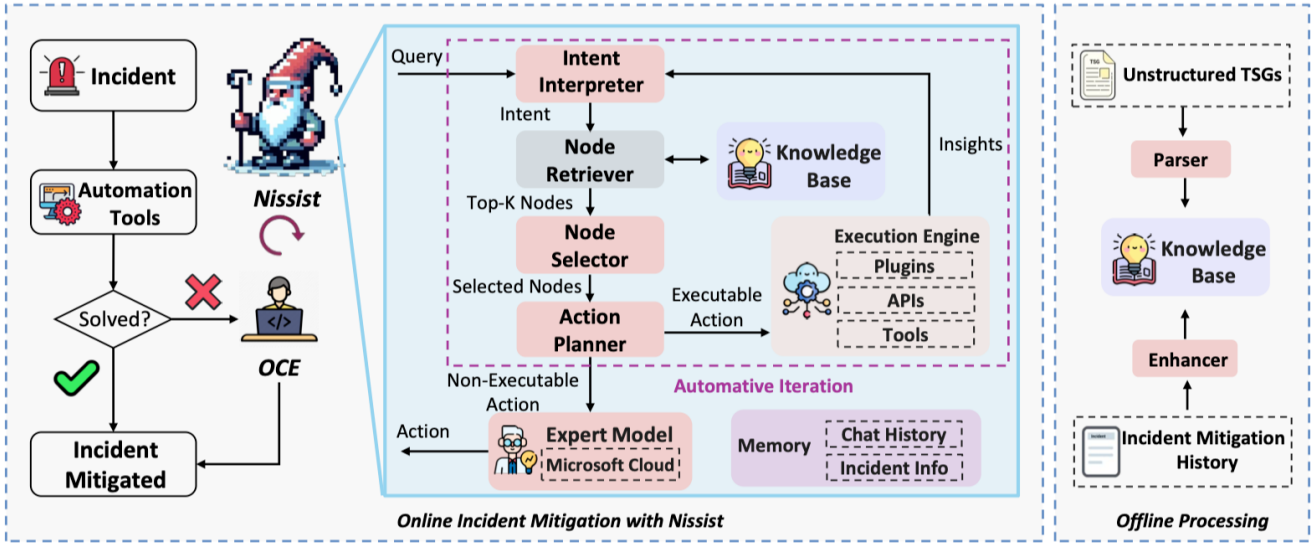
Multi-step Iteration Empowered by Agents



[1] Hamadanian et al (MIT & Microsoft). A Holistic View of AI-driven Network Incident Management. HotNets 2023

Nissist: Incident Mitigation Copilot

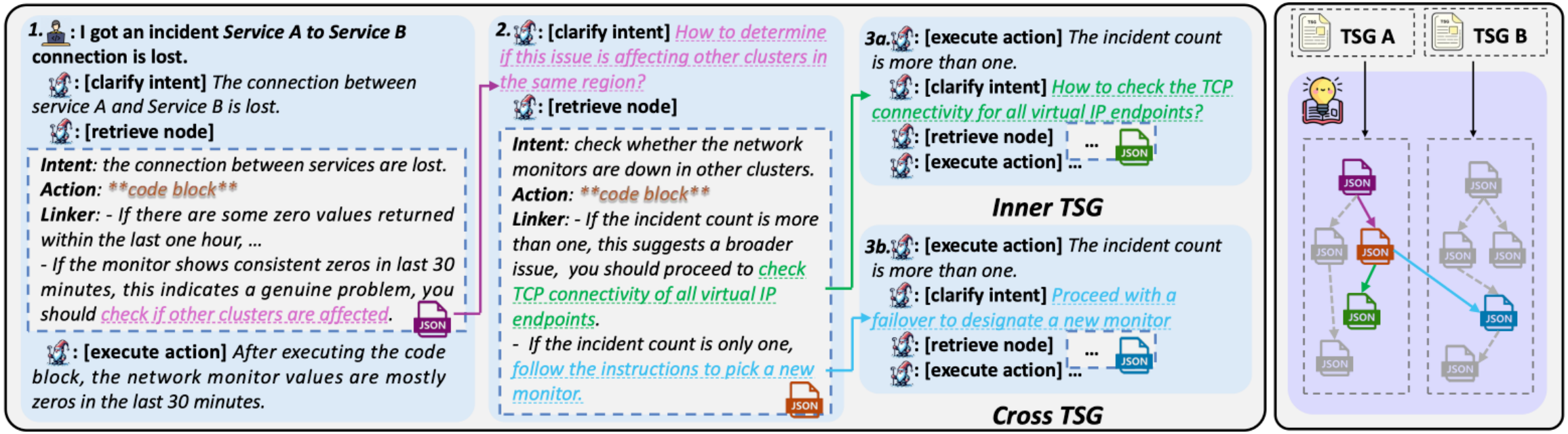
- **Problem and Challenges:**
  - To expedite incident mitigation, troubleshooting knowledge are gathered into Troubleshooting Guides (TSGs).
  - TSGs are often **unstructured and incomplete**, requiring manual interpretation by OCE.
  - TSGs may be **outdated**, lacking the latest incident mitigation knowledge.
- **Approach:**
  - Use LLMs to extract knowledge from unstructured TSGs and incident resolution histories, creating a comprehensive knowledge base.
  - Employ a multi-agent design to enhance the system's ability to accurately identify OCE intent, retrieve relevant information, and provide systematic plans.



An et al (PKU & Microsoft). Nissist: An Incident Mitigation Copilot based on Troubleshooting Guides. arXiv 2024

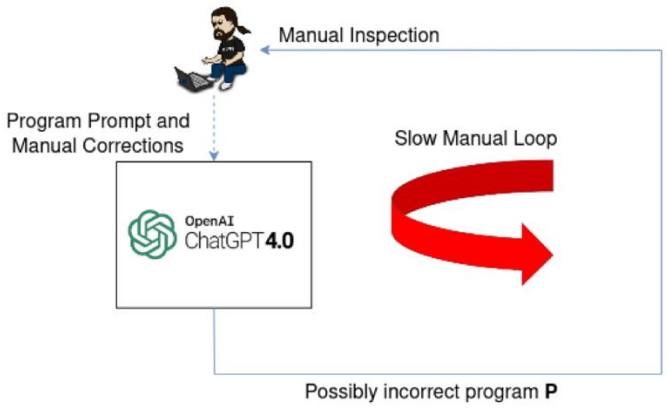
Nissist: Incident Mitigation Copilot

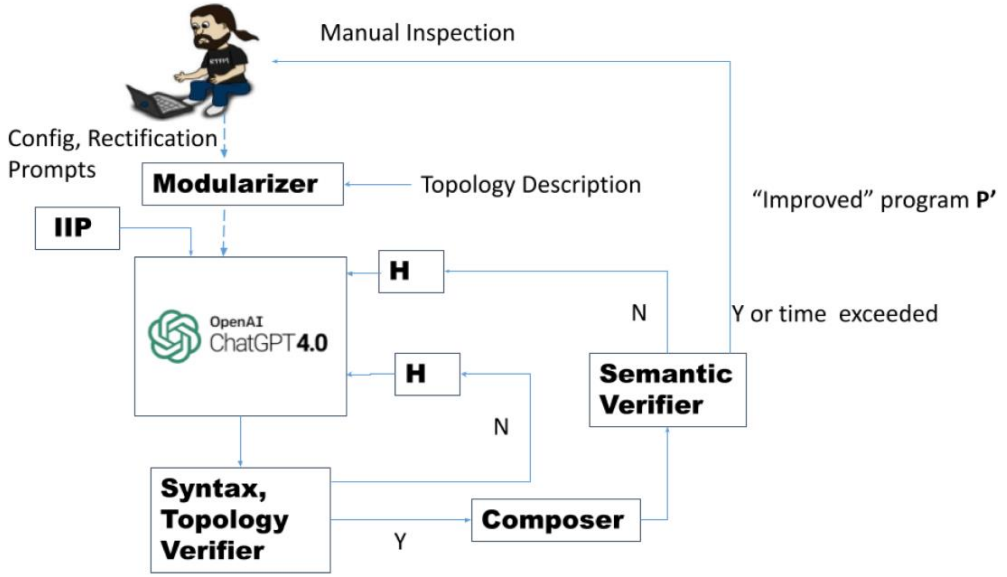
- **Case study**
  - Nissist **correlates the execution outcome with the "Linker"** in the retrieved node, and generates a new intent for the next round of interaction automatically
  - Nissist digests all TSGs into knowledge base, making it possible to discover **connections between nodes located in different TSGs**



Synthesize Correct Router Configurations

- **Challenges and Difficulties**
  - Router configurations are crucial for network operations and are usually written by humans. When using GPT-4 alone, there are significant errors in topology, syntax, and semantics
- **Approach**
  - Verified Prompt Programming (VPP) is proposed, combining GPT-4 with verifiers and using actionable feedback from the verifier to automatically correct errors





Syntax Verifier:
- Batfish

Topology Verifier:
- Custom python code

Semantic Verifier:
- Campion
- Batfish's symbolic route map analysis

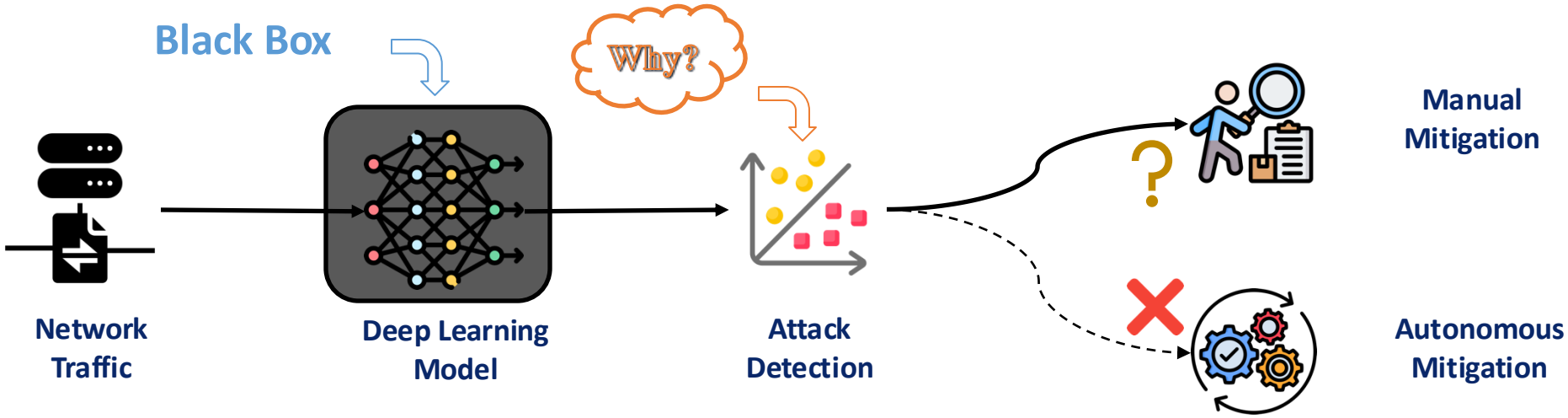
Mondal et al (UCLA). What do LLMs need to Synthesize Correct Router Configurations . HotNets 2023

ShieldGPT： LLM-based Framework for DDoS Mitigation

Weakness of Existing Work

Lack of Explainability

Lack of Mitigation Instructions



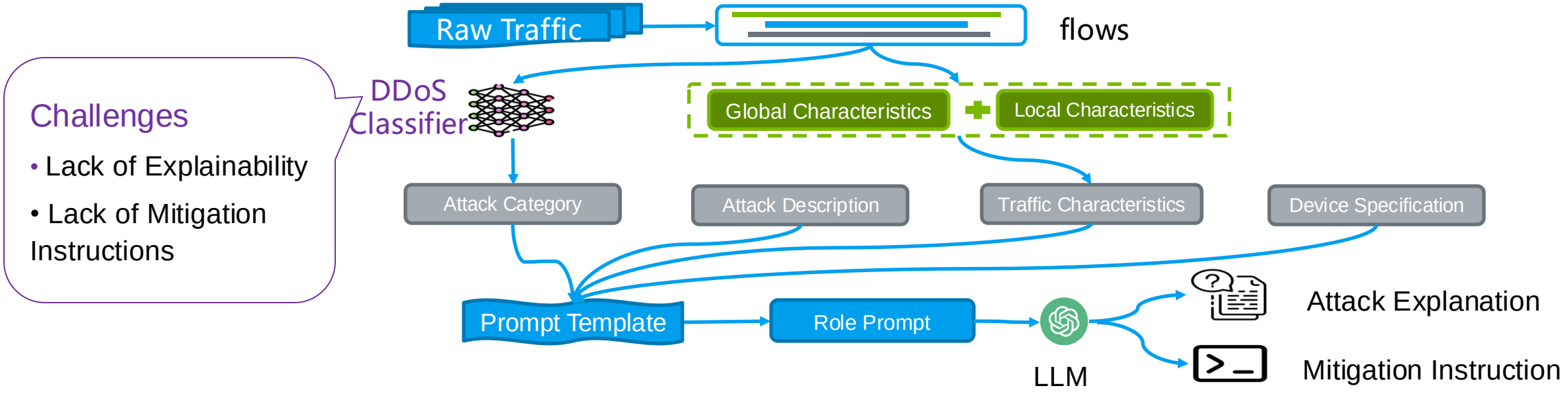
- *Current Explanation: Limited to attribute importance analysis*
- *Black-box nature hinders broader industrial adoption*

- *Cannot provide explanations at a fine-grained level*
- *Automatic generation of mitigation instructions is missing*

Wang et al (THU & Huawei). ShieldGPT: An LLM-based Framework for DDoS Mitigation. APNet 2024

ShieldGPT：LLM-based Framework for DDoS Mitigation

- **Traffic Characterization for Feature Engineering:** using text to describe binary traffic
  - **Global Features:** Statistics of a single flow, reflecting overall traffic behavior.
  - **Local Features:** Key packet fields that reveal interaction details within the traffic.
- **Role Representation for Feature Engineering:** guide LLMs to understand task requirements
  - **Domain Knowledge Injection**: Attack types, attack descriptions, traffic characteristics, device descriptions, etc.



Wang et al (THU & Huawei). ShieldGPT: An LLM-based Framework for DDoS Mitigation. APNet 2024

ShieldGPT： LLM-based Framework for DDoS Mitigation

- **Experiment Results**

Traffic-dependent explanation for a Slowbody attack

1. Large Content-Length Value: The fifth packet has a payload with a Content-Length of 4096, but the actual payload is much smaller. This indicates that the attacker sends less data than expected.
2. Low Packet Rate: The packet rate is 0.644 packets per second, which is relatively low. The attacker declares a large content length but sends very slowly to keep the connection open for as long as possible.
3. Incomplete HTTP POST Requests: The payload includes the header of the POST request but the body of the request is not fully sent.

Traffic-dependent explanation

Table 4: DDoS mitigation instructions formulated for Cisco IOS by ShieldGPT

Instruction	Example Command	Interpretation
ACL	*access-list 101 deny ip <malicious-IP> any*	Block traffic from an identified malicious source IP address
Rate Limiting	*rate-limit input <rate> <normal-size> <burst-size> conform-action transmit exceed-action drop*	Constrain the number of requests a single IP can make within a certain period.
Connection Limiting	*ip http max-connections <number>*	Restrict the number of simultaneous HTTP connections from a single IP address.
TCP Intercept	*ip tcp intercept max-incomplete high <maximum> low <minimum>*	Limit the number of incomplete TCP connections.
Connection Timeout	*ip tcp synwait-time <time>*	Set a short timeout for incomplete TCP connection.
Address Verification	*ip verify unicast source reachable-via rx*	Validate the accessibility of the source IP address in an incoming packet.

Diverse mitigation instructions

The mitigation strategy for a Slowheaders attack

1. **Set Connection Timeout**: Configure the router to close the connection if it does not receive the full HTTP header within a certain time limit.
ip http timeout-policy idle 60 life 86400 requests 10000
2. **Limit the Number of Connections**: Limit the number of simultaneous connections from a single IP address.
ip http max-connections 100

Cisco Router

The mitigation strategy for a Slowbody attack

1. **Limit the number of connections per IP**: The following command will limit the number of connections to 10 per client IP.
iptables -A INPUT -p tcp –syn –dport 80 -m connlimit –connlimit-above 10 -j REJECT
2. **Set the connection timeout**: The following command will set the connection timeout to 30 seconds to allow the server to close idle connections more quickly.
echo 30 > /proc/sys/net/ipv4/tcp_fin_timeout

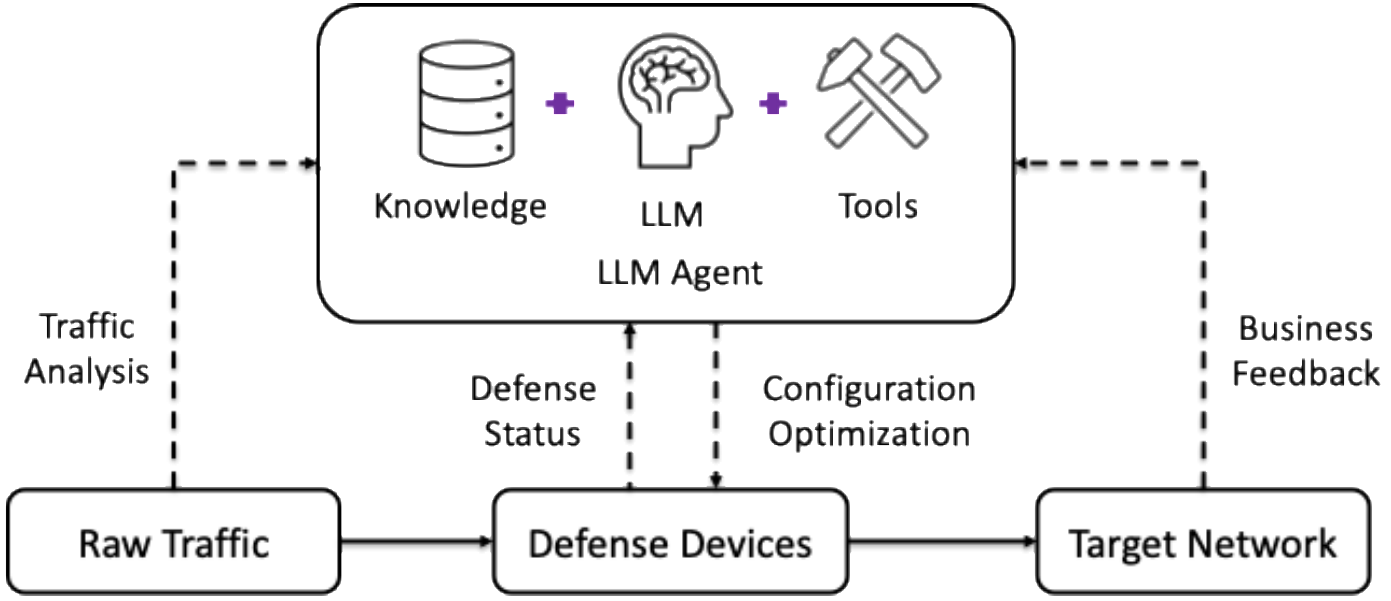
Linux Sever

Executable instructions for different devices

Wang et al (THU & Huawei). ShieldGPT: An LLM-based Framework for DDoS Mitigation. APNet 2024

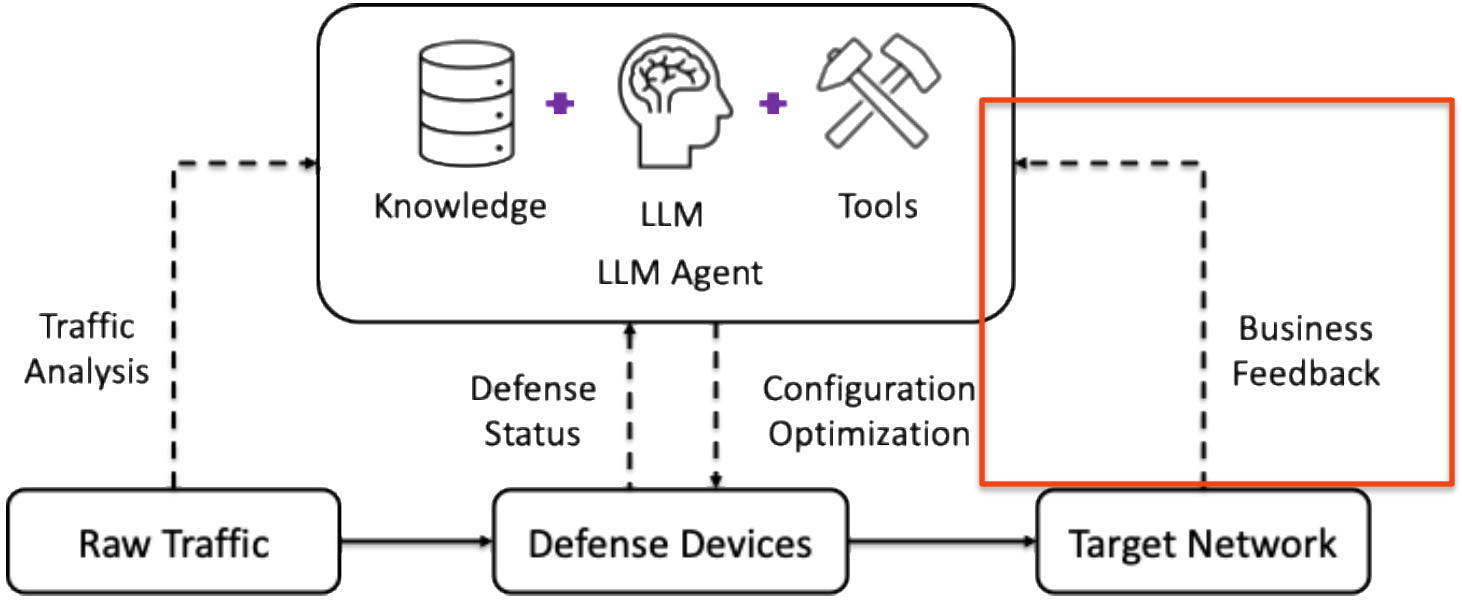
Adaptive Defense Optimization with LLM Agent

- **Problem and Challenges**
  - Cloud providers face growing DDoS threats as more businesses rely on cloud services, increasing the demand for advanced protection capabilities.
    - Traditional ‘default templates + manual adjustments’ struggle to meet the requirements for both reliability and cost-effectiveness when handling intelligent and complex attacks.
- **Approach**
  - The LLM Agent can gain real-time awareness of the attack and defense status and quickly optimize defense strategies based on business feedback



Adaptive Defense Optimization with LLM Agent

- **Problem and Challenges**
  - Cloud providers face growing DDoS threats as more businesses rely on cloud services, increasing the demand for advanced protection capabilities.
    - Traditional ‘default templates + manual adjustments’ struggle to meet the requirements for both reliability and cost-effectiveness when handling intelligent and complex attacks.
- **Approach**
  - The LLM Agent can gain real-time awareness of the attack and defense status and quickly optimize defense strategies based on business feedback



Adaptive Defense Optimization with LLM Agent

- **Preliminary test**
  - **Default defense configuration：** Utilizes The Source IP Traffic Control and The Destination IP Traffic Control policies to limit DDoS attacks, but with **relatively lenient threshold** settings
  - When a DDoS attack occurs, the LLM Agent **perceives a drop in business throughput** and **mistakenly assumes that business traffic is being incorrectly filtered**, leading it to further relax the defense policies

Explanation：

1. Traffic Analysis:

The reduction in business throughput suggests that **some business traffic could be either getting delayed or blocked** inadvertently by the current settings.

2. Reviewing Current Thresholds:

- The sip_tc_max (300 pps) and dip_tc_max (45,000 pps) thresholds appear to have been effective in filtering excessive traffic but might need adjustment to prevent legitimate traffic from being affected.

Policy Adjustments：

```
"UDP": {
        ……
        "sip_tc_enable": "1",
        "sip_tc_unit": "0",
        "sip_tc_max": "350",  // Increase from 300 pps to 350 pps to allow slightly more packets per source IP
        ……
        "dip_tc_enable": "1",
        "dip_tc_unit": "0",
        "dip_tc_max": "50000" // Increase from 45,000 pps to 50,000 pps to ensure higher limit for legitimate peak traffic
    }
```

Adaptive Defense Optimization with LLM Agent

- **Preliminary test**
  - **Default defense configuration：** Utilizes The Source IP Traffic Control and The Destination IP Traffic Control policies to limit DDoS attacks, but with **relatively lenient threshold** settings
  - By **providing more granular feedback**, such as bandwidth utilization and link packet loss rate, the LLM Agent can accurately identify the current defense weaknesses and adjust the configuration in the right direction.

Explanation：

Assessing the Impact of Current Policies:

- The Source IP Traffic Control successfully filtered a significant rate of inbound traffic (18.6 Mbps), indicating its effectiveness but also suggesting that **the limit might be either too lenient or improperly calibrated**.
- The inbound bandwidth is fully utilized, resulting in approximately 8% packet loss. This is problematic as it indicates that **business traffic is likely affected**, thereby reducing business throughput by a similar percentage.

Policy Adjustments：

```
"UDP": {
        ……
        "sip_tc_enable": "1",
        "sip_tc_unit": "0",
        "sip_tc_max": "250",  // Adjusted lower from 300 to a stricter control as attackers may use many IPs
        ……
        "dip_tc_enable": "1",
        "dip_tc_unit": "0",
        "dip_tc_max": "50000"  // Incrementally adjusted to allow more traffic and observe if legitimate traffic improves
    }
```

Challenges and Future Directions

- **Intelligent Planning**
  - Essential for accomplishing complex tasks that involve long-term objectives and multi-step decision-making
  - A promising direction is the development of LLM-based agents that integrate **perception**, **decision-making**, and **action** capabilities.
    - Enhanced situation assessment
    - Effective strategies based on feedback
    - Facilitates rapid iteration
- **Understanding Multimodal Data**
  - Multimodal data plays a critical role in the networking domain.
  - Potential Solutions:
    - Achieve better **model alignment** and enhance understanding and reasoning capabilities through targeted **fine-tuning**.
    - Incorporate diverse multimodal datasets in the **pre-training** phase and design multi-objective tasks to strengthen pre-training effectiveness.

Challenges and Future Directions

- **Prompt Engineering for Deliberate Reasoning**
  - Many networking tasks involve integrating multiple intermediate results to reach a final conclusion
  - Various methods have been explored to facilitate **multi-path reasoning processes**
    - Tree of Thoughts (ToT)
    - Graph of Thoughts (GoT)
  - Design effective **reflection** prompts, enabling LLM to learn from failures and generate more accurate results
- **Network-specific LLMs**
  - Construct LLMs specialized for the networking domain to enhance efficiency and performance
  - Utilize diverse network data from various scenarios to establish a foundational model, followed by fine-tuning for specific tasks
  - Must be **lightweight and efficient** to accommodate temporal and spatial constraints in network environments
    - Knowledge distillation，model quantization，novel frameworks like Mamba

Challenges and Future Directions

- **Autonomous Tools Utilization**
  - Few-shot learning leads to incomplete comprehension and suboptimal performance.
  - Fine-tuning incurs high overhead and reduces flexibility
  - **Represent tools as tokens**, enabling LLMs to learn their embedded representations
  - Generate tool tokens as needed, allowing the model to switch between reasoning mode and tool mode and maximize the ability to handle both natural language and tool-specific tasks
- **Validation Environment**
  - Ensuring the reliability and safety of applying LLMs in networking presents a critical challenge
  - Integrate LLMs with validation environments, such as **digital twins**
    - Simulate system behavior, operate synchronously with actual systems, monitor system status, and provide real-time feedback
    - Enables continuous adjustments and optimization of LLMs' output, leading to effective iteration

Thanks

Xiaohui Xie
Tsinghua University
xiexiaohui@tsinghua.edu.cn