draft-cui-nmrg-llm-net-00 IETF 122 NMRG

A Framework for LLM-Assisted Network Management with Human-in-the-Loop

Yong Cui, **Mingzhe Xing**, Lei Zhang

Tsinghua University, Beijing Zhongguancun Laboratory

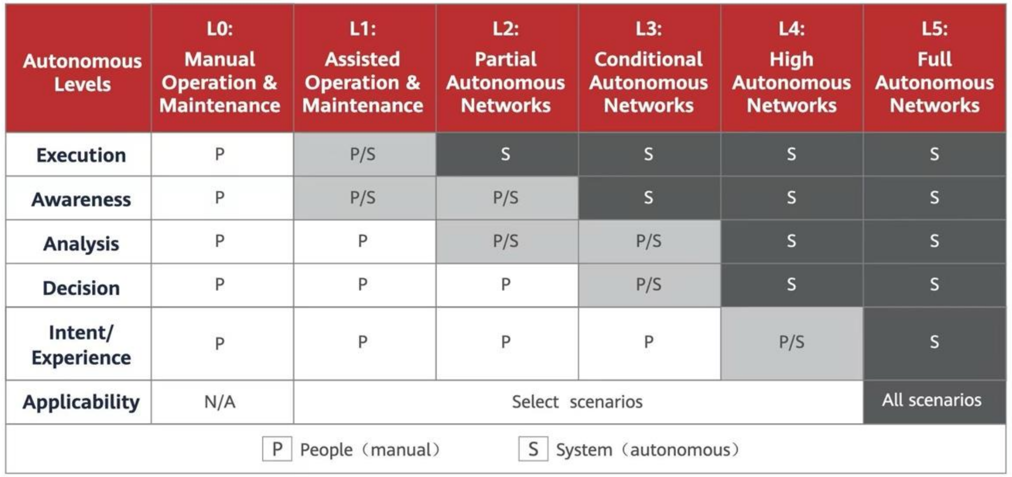
Mar, 2025

Outline

- Motivation
- Framework Overview
- Data Model
- Security Consideration
- Future Work

Motivation

- Challenges in Traditional Network Management
  - Complex and dynamic network environment
  - Diverse intents and demands
  - Rapid evolution and iteration
  - Learning curve on vendor-specific devices

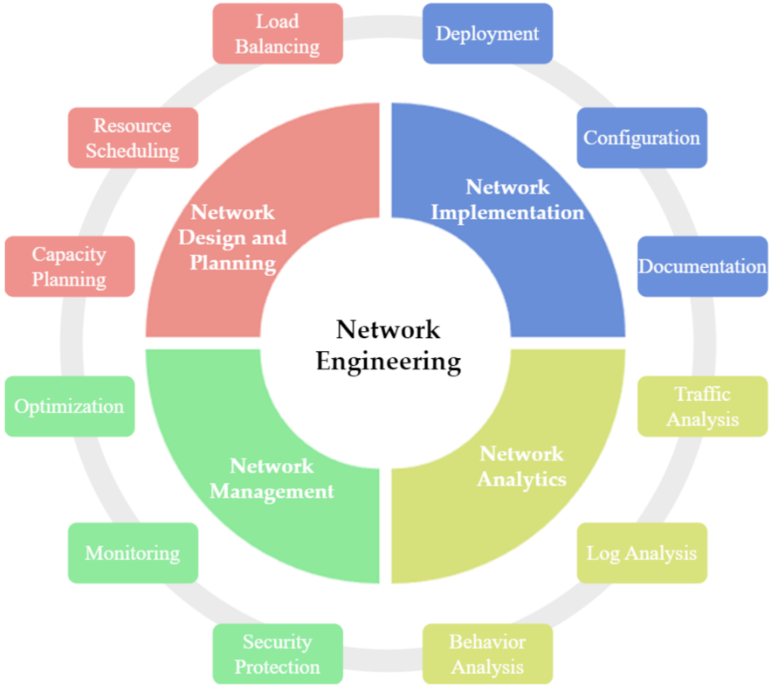


Autonomous Levels	L0: Manual Operation & Maintenance	L1: Assisted Operation & Maintenance	L2: Partial Autonomous Networks	L3: Conditional Autonomous Networks	L4: High Autonomous Networks	L5: Full Autonomous Networks
Execution	P	P/S	S	S	S	S
Awareness	P	P/S	P/S	S	S	S
Analysis	P	P	P/S	P/S	S	S
Decision	P	P	P	P/S	S	S
Intent/ Experience	P	P	P	P	P/S	S
Applicability	N/A	Select scenarios				All scenarios
P People（manual） S System（autonomous）						

Table 1-1 Levels of Autonomous Networks

- Vision: Autonomous Network Management [TM-IG1230]
  - ✓ Zero-X: zero wait, zero touch, and zero trouble
  - ✓ Self-X: self-configuration, self-healing, self-optimizing and self-evolving

Revolutionizing NM with LLM



- ➢ Opportunity: Large Language Models
  - ➢ Multi-modal data understanding
  - ➢ Logical reasoning
  - ➢ Generalization

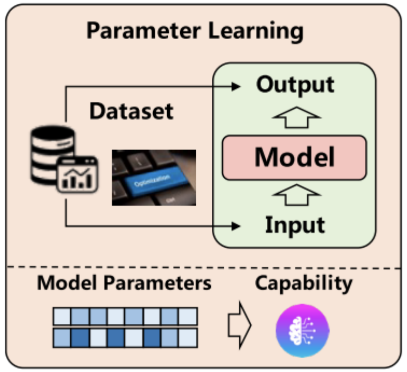
Based on Assumptions

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

R： Resolution
D： Delay
C： Change rate

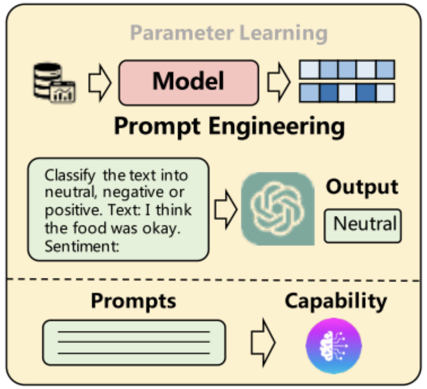
Mathematical Modeling

Domain dataset



Machine Learning

Extensive datasets and Large-scale parameters



Classify the text into neutral, negative or positive. Text: I think the food was okay. Sentiment:

LLM

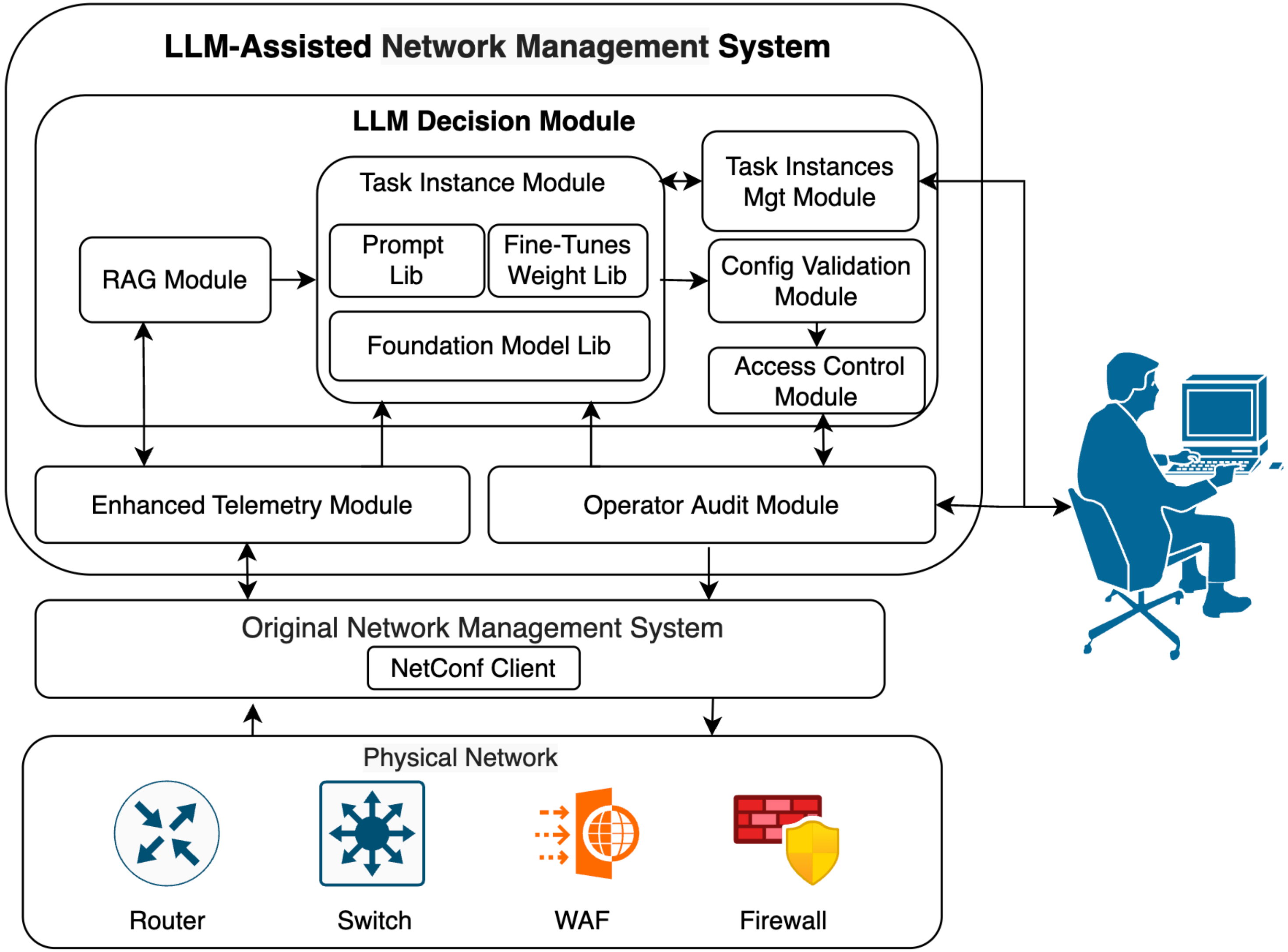
Increasing Attention

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
KTH Royal Institute of Technology	NetConfEval: Can LLMs Facilitate Network Configuration?	CoNEXT 24
Microsoft & UIUC	Automatic Root Cause Analysis via Large Language Models for Cloud Incidents	EuroSys 24

Human Still “in the Loop”

- **Consensus** in NMRG Charter:
  - There will be intermediate levels where the **human users remain “in the loop”** and are **progressively assisted and replaced** by more and more intelligent mechanisms
  - **Interfaces between humans and a self-driving system** are important and required to allow bidirectional communications
- **LLM-Assisted** Network Management Framework with **Human-in-the-Loop**
  - The **framework** components that build the intelligent autonomous system
  - The **workflow** of autonomous decision and “Human in the Loop”
  - The **interface** of human operator and LLM-assisted system

Framework Overview



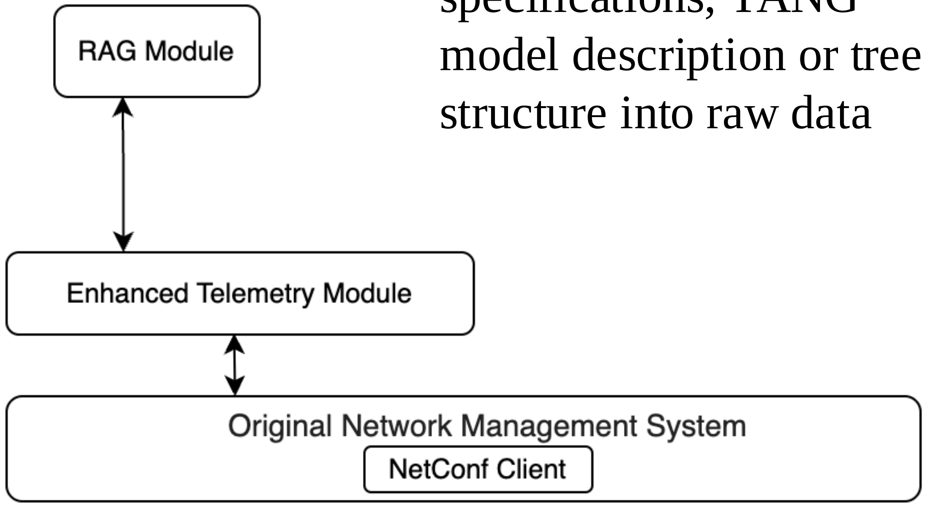
Key Components:

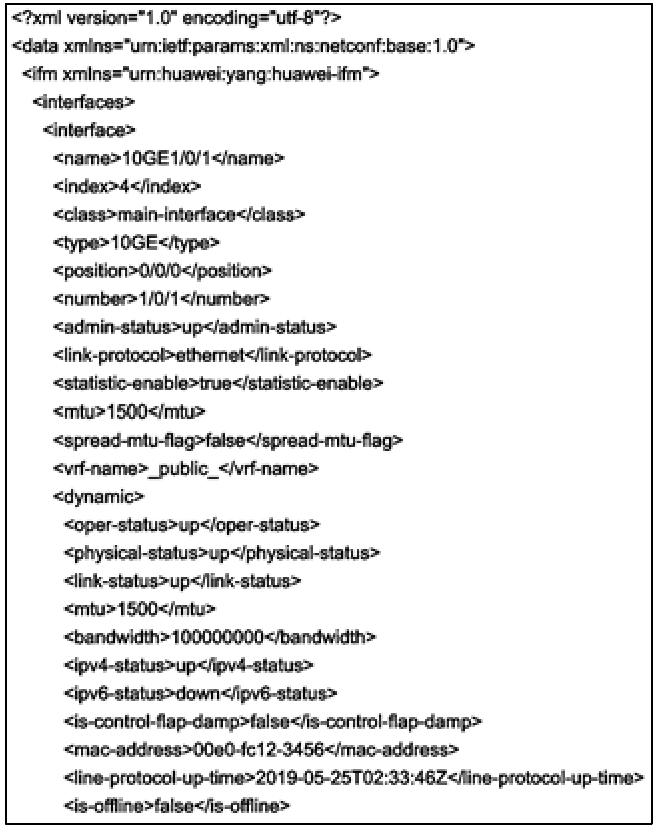
- **Enhanced telemetry module** improves the semantics of raw telemetry data
- **LLM-decision module** specifies a task instance to generate configurations
- **Human operator audits** the configuration passed by the validation and access control modules

Enhanced Telemetry Module

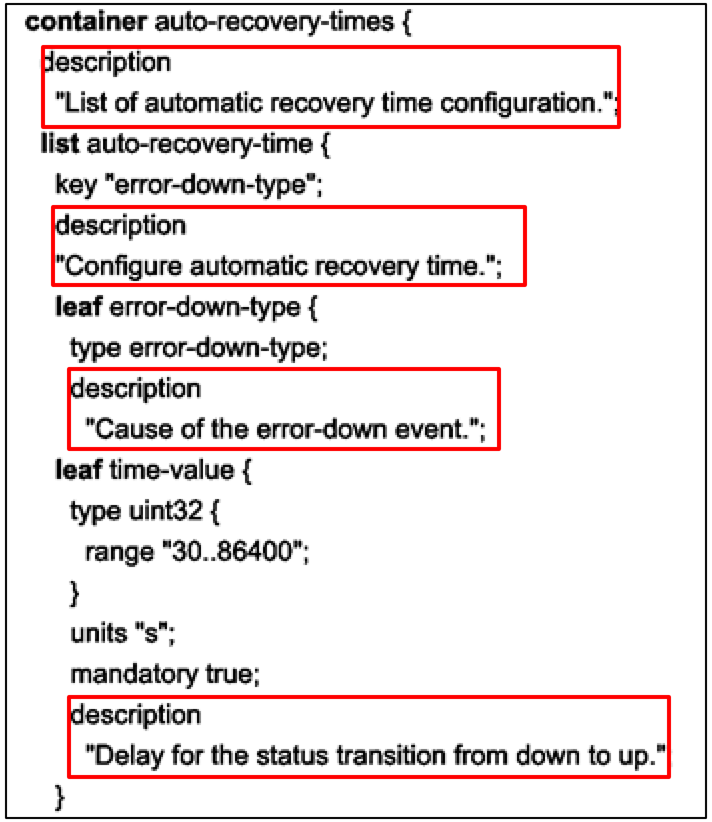
- ➢ Telemetry data retrieved via NETCONF, e.g., in XML format, lacks field descriptions and structured metadata
  - ➢ Pretrained LLMs lack this contextual knowledge, and can lead to misinterpretation and erroneous reasoning

Solution: Inject device specifications, YANG model description or tree structure into raw data





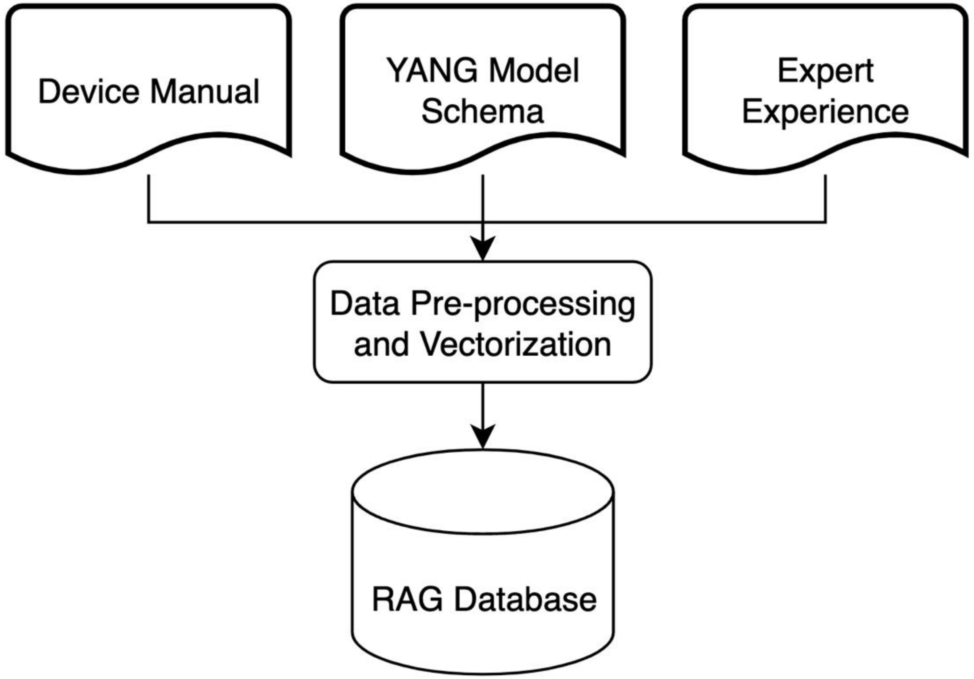
```
<?xml version="1.0" encoding="utf-8"?>
<data xmlns="urn:ietf:params:xml:ns:netconf:base:1.0">
 <ifm xmlns="urn:huawei:yang:huawei-ifm">
  <interfaces>
   <interface>
    <name>10GE1/0/1</name>
    <index>4</index>
    <class>main-interface</class>
    <type>10GE</type>
    <position>0/0/0</position>
    <number>1/0/1</number>
    <admin-status>up</admin-status>
    <link-protocol>ethernet</link-protocol>
    <statistic-enable>true</statistic-enable>
    <mtu>1500</mtu>
    <spread-mtu-flag>false</spread-mtu-flag>
    <vrf-name>_public_</vrf-name>
    <dynamic>
     <oper-status>up</oper-status>
     <physical-status>up</physical-status>
     <link-status>up</link-status>
     <mtu>1500</mtu>
     <bandwidth>100000000</bandwidth>
     <ipv4-status>up</ipv4-status>
     <ipv6-status>down</ipv6-status>
     <is-control-flap-damp>false</is-control-flap-damp>
     <mac-address>00e0-fc12-3456</mac-address>
     <line-protocol-up-time>2019-05-25T02:33:46Z</line-protocol-up-time>
     <is-offline>false</is-offline>
```



```
container auto-recovery-times {
 description
  "List of automatic recovery time configuration.";
 list auto-recovery-time {
  key "error-down-type";
  description
  "Configure automatic recovery time.";
  leaf error-down-type {
   type error-down-type;
   description
    "Cause of the error-down event.";
  leaf time-value {
   type uint32 {
    range "30..86400";
   }
   units "s";
   mandatory true;
   description
    "Delay for the status transition from down to up."
  }
```

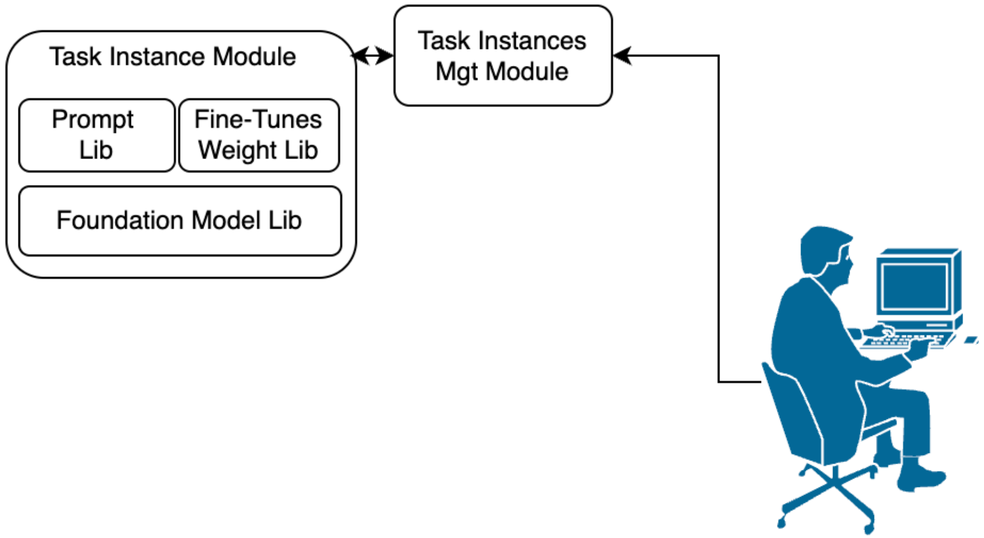
LLM Decision Module

- Retrieve-Augmented Generation (RAG) Module
  - Data Source: device documentation, expert knowledge, and YANG model schemas
  - Knowledge Compensate: retrieve relevant knowledge by text or vector similarity



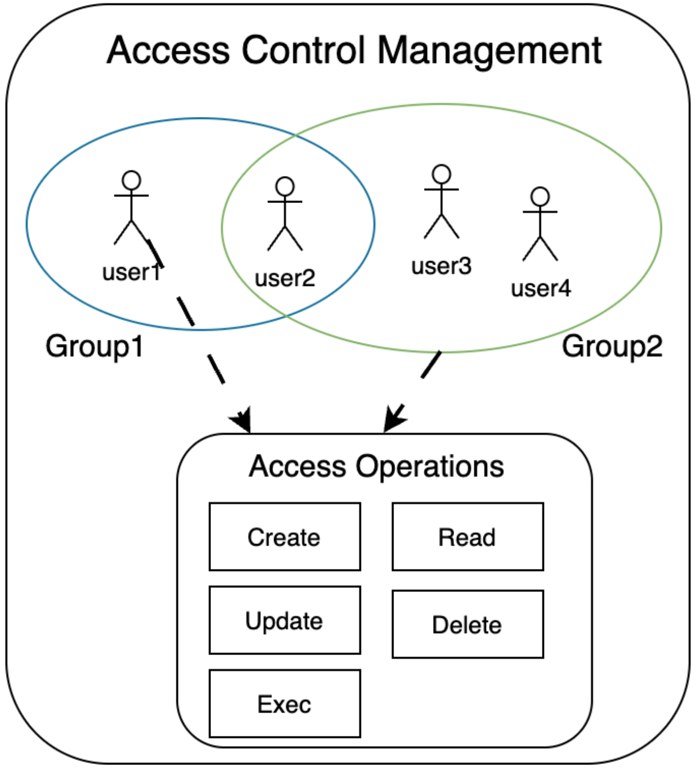
LLM Decision Module

- Task Instance Management
  - Operator specifies a task, e.g., traffic analysis, traffic optimization, or attack mitigation
  - Task instance management module creates a task instance
    - Foundation Model Selection (e.g., GPT-4, LLaMA, and DeepSeek)
    - Prompt Selection (define the task description, and input and output formats)
    - Fine-Tune Weight Selection (adaptation weights trained on private datasets)



LLM Decision Module

- Access Control Module
  - **RFC8341: Network Configuration Access Control Model**
  - User and Group
    - Each task instance should be registered as a specific user
    - A group consists of zero or more members, and a task instance can belong to multiple groups
  - Access Operation Types
    - create, read, update, delete, and execute
  - Action Types
    - permit or deny
  - Rule List



Simplified abstraction of RFC8341

Operator Audit Module

- **LLM-Assisted NM System → Human Operator**

  - □ Generated Network Configuration
  - □ Confidence Score
  - □ Error Message

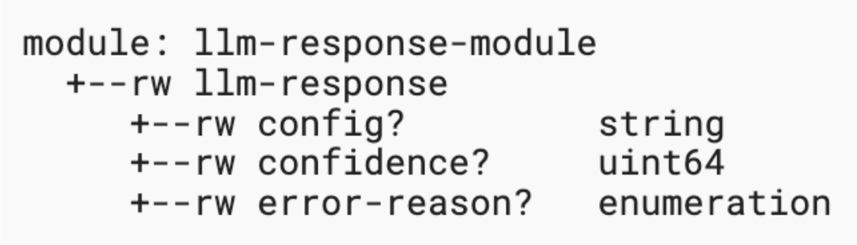
- **Human Operator → LLM-Assisted NM System**

  - ✓ Result Verification
  - ✓ Compliance Check
  - ✓ Security Verification
  - ✓ Suggestions and Corrections

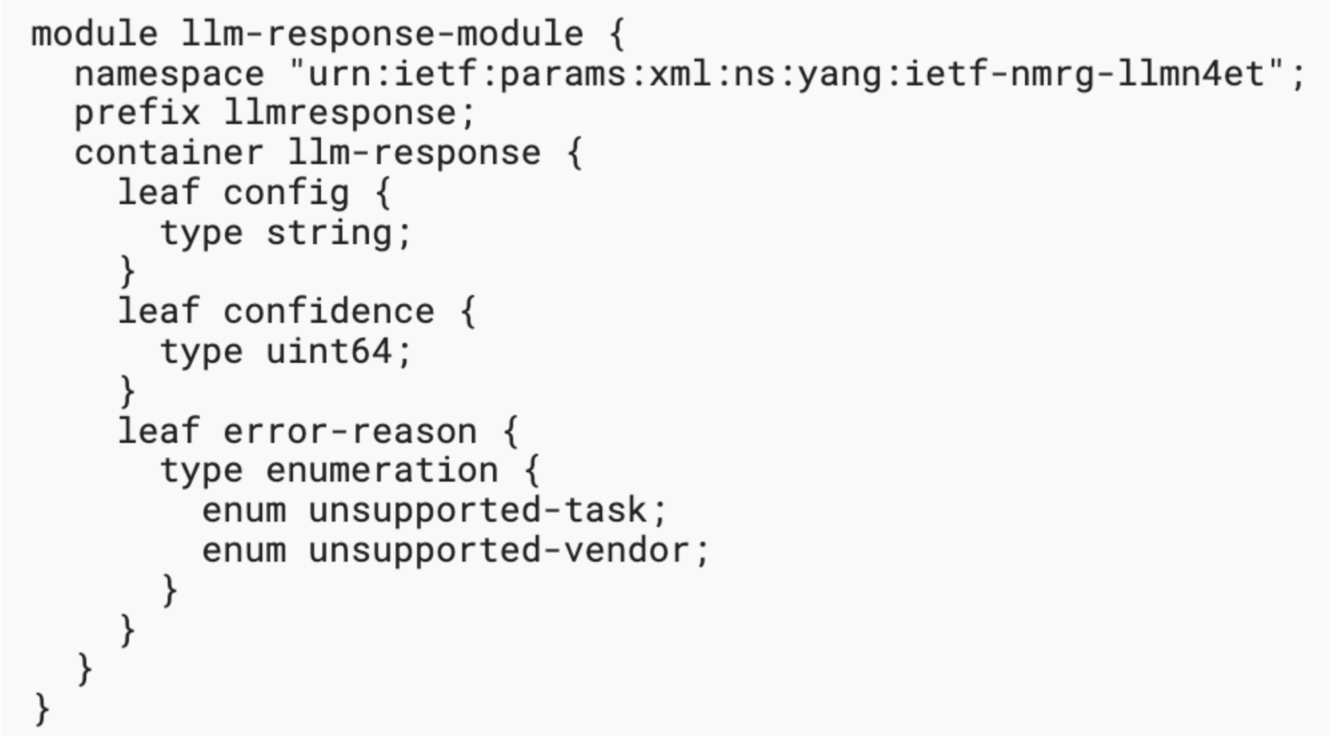
  - □ Audit Timestamp of the audit action
  - □ LLM Task Instance ID
  - □ Operator decisions (approval, rejection, modification, or pending)
  - □ Final executed command

Data Models

- **LLM-Assisted NM System → Human Operator**



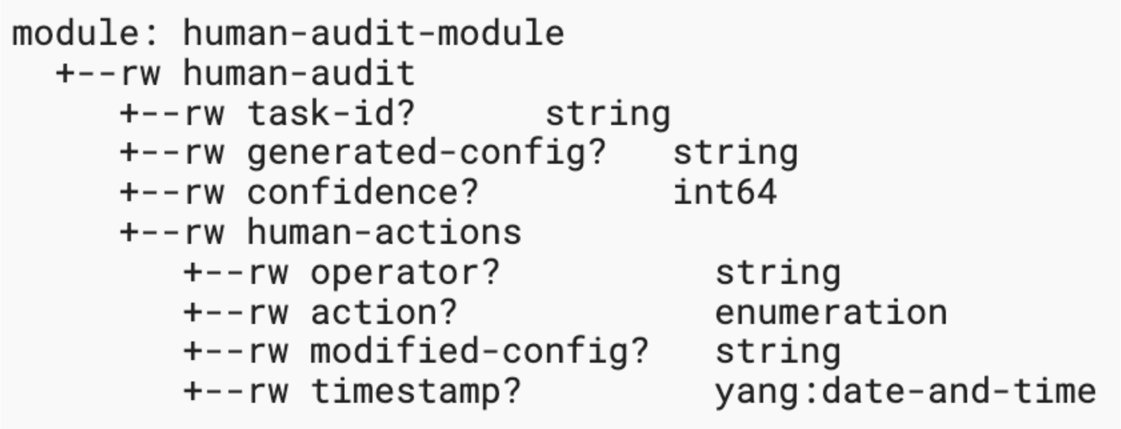
```
module: llm-response-module
  +--rw llm-response
     +--rw config?         string
     +--rw confidence?     uint64
     +--rw error-reason?   enumeration
```



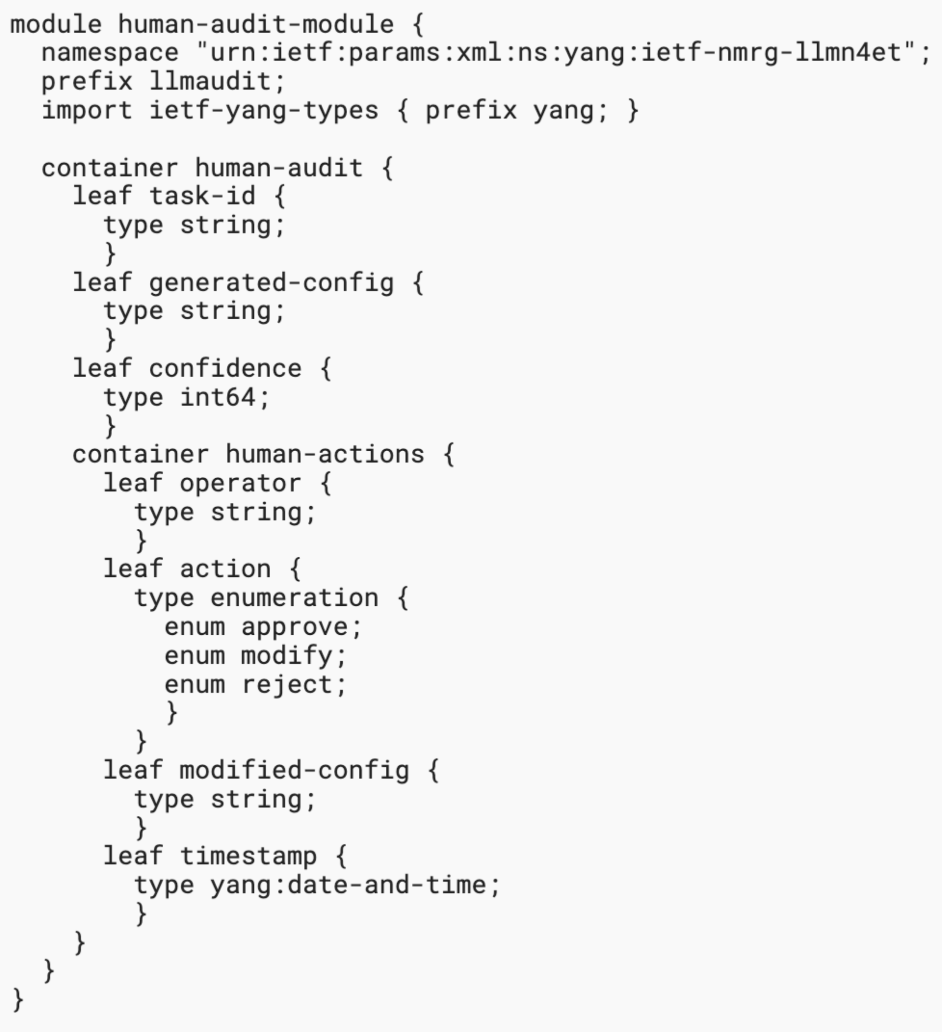
```
module llm-response-module {
  namespace "urn:ietf:params:xml:ns:yang:ietf-nmrg-llmn4et";
  prefix llmresponse;
  container llm-response {
    leaf config {
      type string;
    }
    leaf confidence {
      type uint64;
    }
    leaf error-reason {
      type enumeration {
        enum unsupported-task;
        enum unsupported-vendor;
      }
    }
  }
}
```

Data Models

- **Human Operator → LLM-Assisted NM System**



```
module: human-audit-module
  +--rw human-audit
     +--rw task-id?           string
     +--rw generated-config?      string
     +--rw confidence?            int64
     +--rw human-actions
        +--rw operator?             string
        +--rw action?               enumeration
        +--rw modified-config?      string
        +--rw timestamp?            yang:date-and-time
```



```
module human-audit-module {
  namespace "urn:ietf:params:xml:ns:yang:ietf-nmrg-llmn4et";
  prefix llmaudit;
  import ietf-yang-types { prefix yang; }

  container human-audit {
    leaf task-id {
      type string;
      }
    leaf generated-config {
      type string;
      }
    leaf confidence {
      type int64;
      }
    container human-actions {
      leaf operator {
        type string;
        }
      leaf action {
        type enumeration {
          enum approve;
          enum modify;
          enum reject;
          }
        }
      leaf modified-config {
        type string;
        }
      leaf timestamp {
        type yang:date-and-time;
        }
    }
  }
}
```

Future Work

- Security Considerations
  - Model Hallucination: LLMs may produce malformed or invalid configurations
  - Training Data Poisoning: LLMs trained on malicious or biased data could exhibit unintended behavior or introduce security vulnerabilities
- Future Work
  - Define the task instance management interface
  - Detail the audit process of human operator
  - Integrate the Intent Based Network (IBN) into the audit interface

Thanks!

We welcome collaborators!

Q&A

Mingzhe Xing
xingmz@zgclab.edu.cn