From Fragmented Benchmarks to an Interoperable Format for Evaluating Network LLM Agents

Chenguang Du

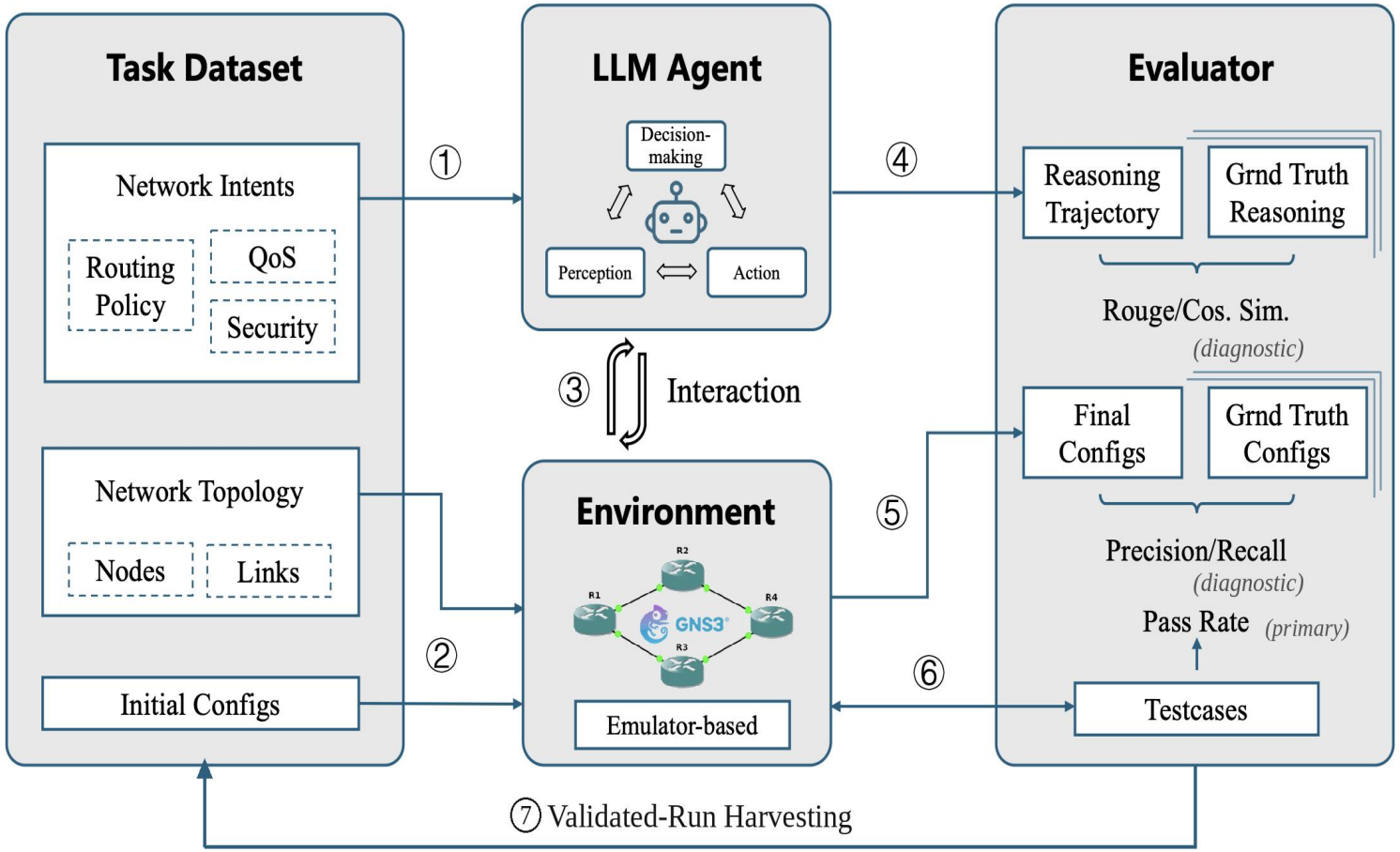
Beijing Zhongguancun Laboratory

July, 2026

Motivation: Fragmented Benchmarks

- Challenges
  - A growing number of projects are building benchmarks for AI agents in networking
  - Their task formats, testcases, validation methods, and evaluation metrics are incompatible
  - Scenarios are difficult to reuse, and results are difficult to compare
- Our proposal: draft-cui-nmrg-llm-benchmark
- A common, extensible format for network-agent evaluation
  - Shared task schema: **intent, topology, initial state, valid solutions, executable testcases**
  - An emulator automation layer (Harness) executes tasks end-to-end

Framework Overview



- Key Components:
  - Task Dataset
  - Emulator Environment
  - LLM Agent
  - Evaluator
- Steps:
  - ① Task Assignment
  - ② Environment Setup
  - ③ **Emulator-based Interactive Execution**
  - ④ Reasoning Export
  - ⑤ Configuration Export
  - ⑥ Scoring
  - ⑦ **Passed test case become new ground-truth solutions**

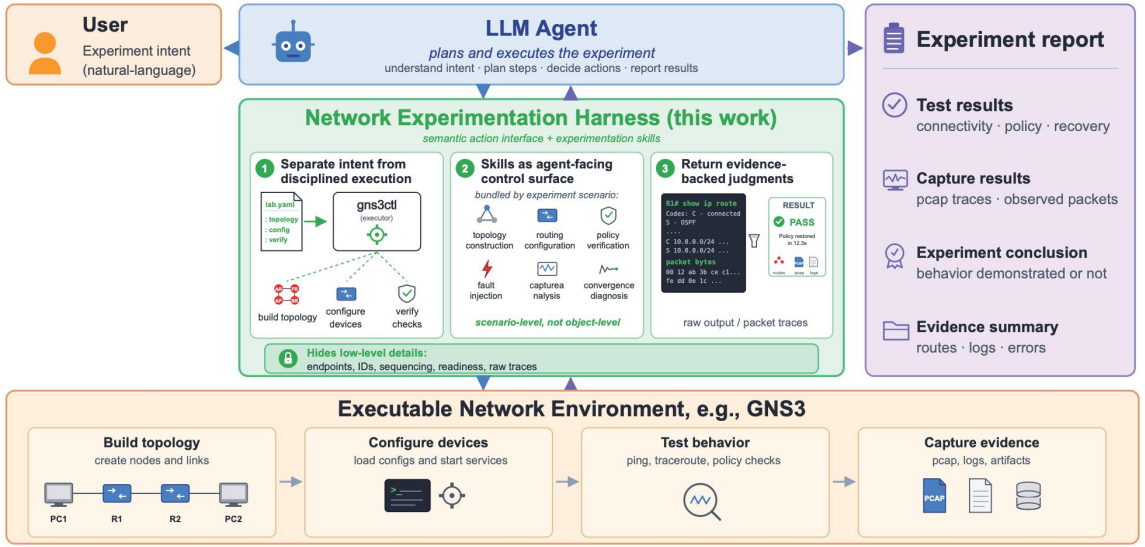
Execution is interactive in the emulated environment
The results feed back into the dataset

Task Schema as an Interoperable Exchange Format

- **One task = a JSON object, with five parts:**
  - **intent · topology · initial configs · ground-truth solutions · testcases**
- New in draft-02: a task can have multiple solutions
  - each solution carries its own configs and reasoning trace
- Existing datasets can be imported; Our tasks can be exported to other benchmarks
- **The community needs a common exchange format**

A baseline for the benchmark

- AI agent waste many steps on sequencing, timing, and verification
- Network Experiment Harness = **semantic action interface + reusable experiment skills**
  - It enables agents to complete networking tasks end-to-end.
  - it builds the topology, configures the devices, runs the tests, and reports results
- Efficiency (4 tasks, avg.)
  - **vs. Raw API : wall-clock time ↓47% , token use ↓81% ; lead widens with task complexity**



Thanks!

We welcome collaborators!

Draft: datatracker.ietf.org/doc/draft-cui-nmrg-llm-benchmark
Code: github.com/AutoNetResearch/GNS3-MCP-Adapter

Q&A