Evaluating Agentic Configuration Repair for Computer Networks

Rufat Asadli[1], Benjamin Hoffman*[1], Ioannis Protogeros*[1], Laurent Vanbever[1]
[1]Networked Systems Group, D-ITET, ETH Zurich

*Equal contribution, order determined by coin flip.

Problem Statement

Misconfigurations in computer networks cause major Internet outages, and even **frontier LLMs struggle** to repair them.


The New York Times

Gone in Minutes, Out for Hours: Outage Shakes Facebook

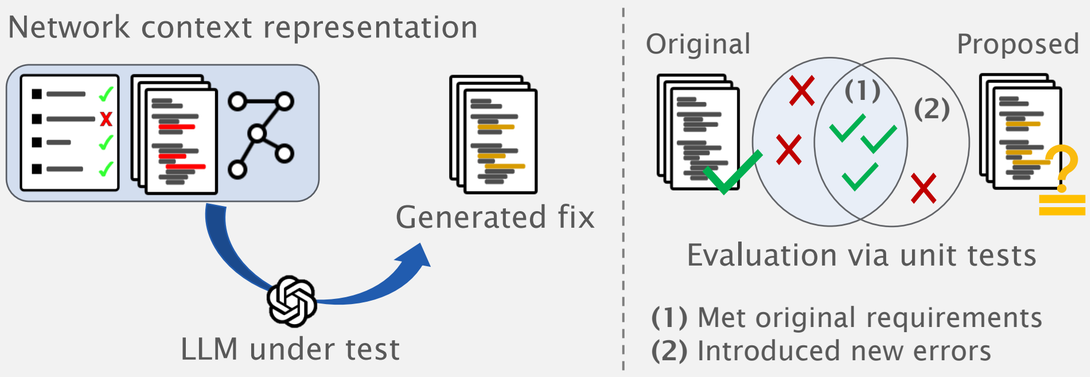
When apps used by billions of people worldwide blinked out, lives were disrupted, businesses were cut off from customers —

Challenges: Real-world network data is **vast** and **noisy**, and edits on network state have hard-to-predict topology-wide effects.

This work: We built a minimal **agentic scaffold** that improves base LLMs with **+12% efficacy** and **−17% safety regressions**.

Benchmark & Methodology

Task definition: We evaluate agents on Cornetto, a benchmark for LLMs in *network configuration repair* (Protogeros et al., 2026).

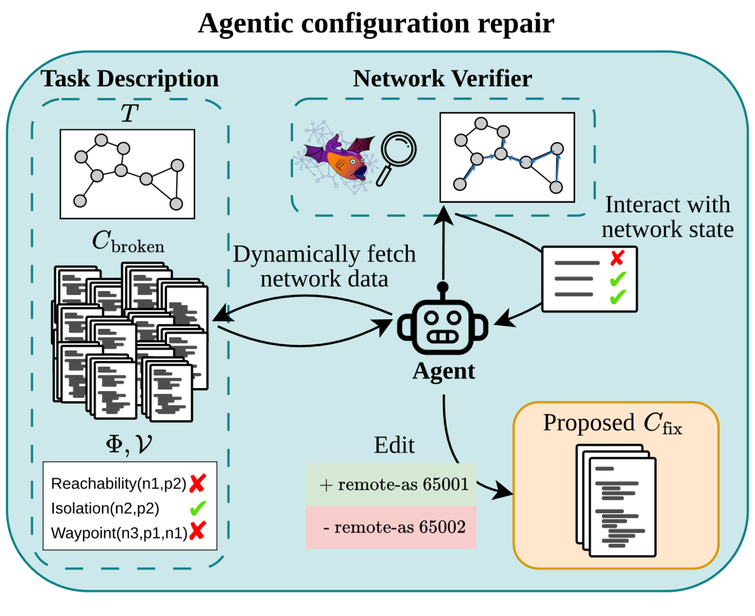


Dataset: 231 scenarios with 27 highly-disruptive adversary types.

Evaluation: We measure the *Fix Score* and *Regression Rate,* as for **(1)** satisfied and **(2)** failed **post-fix** unit tests, respectively.

Dynamic Agentic Scaffolding

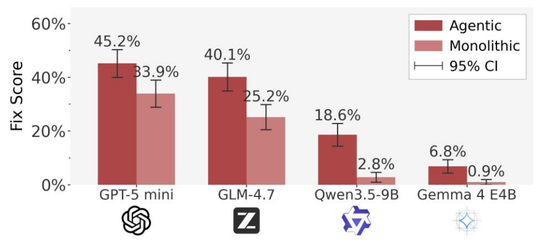
We proposed a **multi-turn agent** with custom tools: (1) *verifier feedback*, (2) *context retrieval*, (3) *search-and-replace repair*.



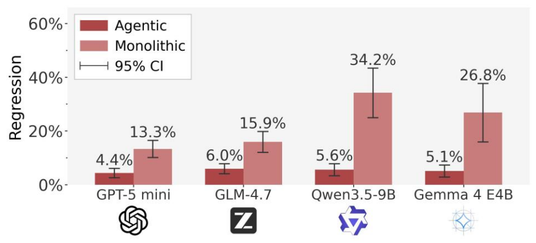
Preliminary Evaluation

Main Leaderboard

Across four open- and closed-source LLMs, agentic scaffolding consistently **outperforms** monolithic prompting.



(a) Fix Score (↑ *is better*)



(b) Regression Rate (↓ *is better*)

Open-source models see fix scores **rise up to 7×** and regression rates drop **from 34% to under 6%**.

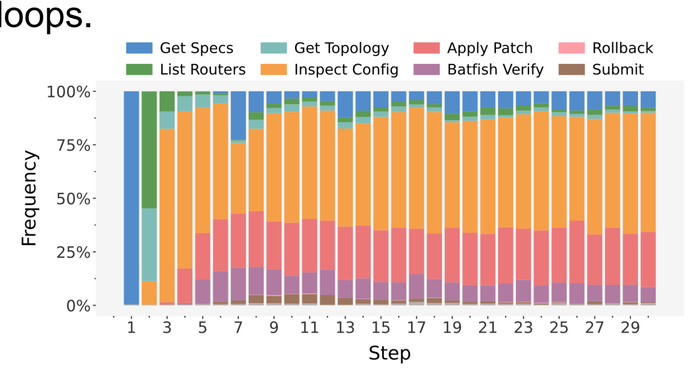
Tool Importance

- **Verifier feedback** cuts safety regressions but can make strong models conservative.
- **Context retrieval** helps small models, while long-context models perform better with prefilled inputs.

Component	Setting	Fix ↑	Regr. ↓
Verifier Feedback			
GPT-5 MINI	Without feedback	**45.2**	4.4
	With feedback	40.8	**1.3**
	Δ	-4.4	-3.1
QWEN3.5-9B	Without feedback	18.6	5.6
	With feedback	**20.0**	**2.9**
	Δ	+1.4	-2.7
Context Retrieval			
GPT-5 MINI	Prefilled retrieval	**49.1**	1.3
	Dynamic retrieval	40.8	1.3
	Δ	-8.3	0.0
QWEN3.5-9B	Prefilled retrieval	14.8	8.3
	Dynamic retrieval	**20.0**	**2.9**
	Δ	+5.2	-5.4

Agent Trajectory

Agents follow a consistent workflow: file analysis first, then patch-and-verify loops.



Limitations and Future Work

Our agentic setup improves base LLMs' efficacy and safety, but key directions remain:

- Richer **LLM–network interfaces** for real-time applications.
- Stronger **safety guardrails**, with human-in-the-loop approval.
- **RL with process-level rewards** for small models as efficient, privacy-preserving alternatives to proprietary ones.