NUS | Computing
National University of Singapore

GroundLM Workshop

EMNLP2026 BUDAPEST

Graph Evidence Is Not Text Evidence

Serialization Faithfulness as Structural Invariance in Graph-Grounded Language Models

Lifan Hu National University of Singapore
lifan.hu@u.nus.edu

Code & evaluation

github.com/Anormalm/graph_evidence

A model can answer graph questions correctly and still change its answer when only the node names change.

98.9%
answer accuracy (Best)

→ **72.9%**
of its correct answers survive renaming

27.1% of its correct answers change

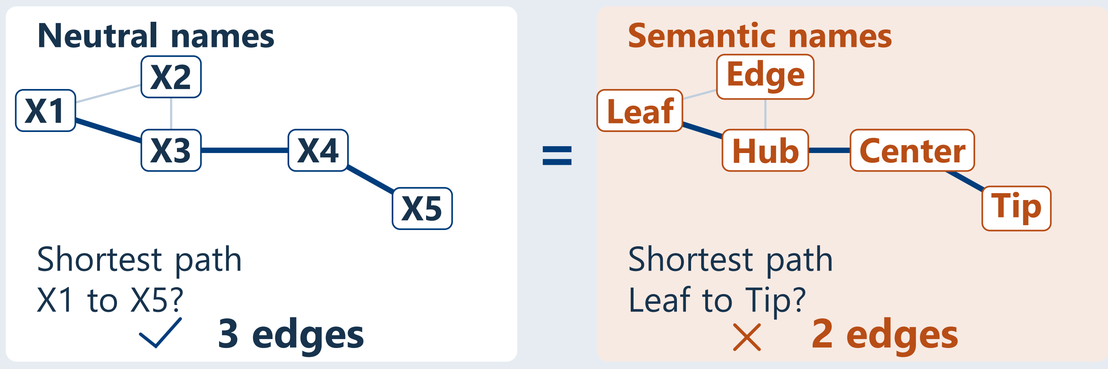
01 Principle

LLMs read serialized text. Faithful graph reasoning should survive changes to node names.

$$(G, q) \sim (G', q') \implies F_\theta(z(G, q)) = F_\theta(z(G', q'))$$

Why? Relabeling preserves adjacency, paths, degrees and motifs. The answer should stay the same.

02 Rename & re-ask



03 Results across models

Model	Accuracy ↑	1 − FS ↓	OS ↓	Δ_{sem} ↓	1 − DR ↓	CF_M ↑
DeepSeek-V4-Flash	**98.9**	**2.5**	**1.8**	26.8	**2.6**	**72.9**
gpt-5.4-mini	93.6	8.4	4.6	15.2	6.6	53.3
Llama-4-Scout	68.9	25.9	14.2	9.1	30.6	65.8
Qwen3.6-35B-A3B	54.1	32.5	8.6	**6.1**	31.8	57.5

Values in %; Δ_{sem} in points. Darker = worse within each column; **bold = best**.
FS format stability; OS order sensitivity; DR distractor robustness.
CF_M: share of correct answers that stay correct after semantic renaming.

04 Findings

Targeted interference
DeepSeek consistency: 96.9% to 70.1% under semantic names; other audits stay strong. gpt-5.4-mini keeps 53.3% of correct answers.

Broad instability
Qwen and Llama fail across audits. Their small gaps (6.1 / 9.1 points) reflect low neutral-name consistency.

Measure invariance next to accuracy.

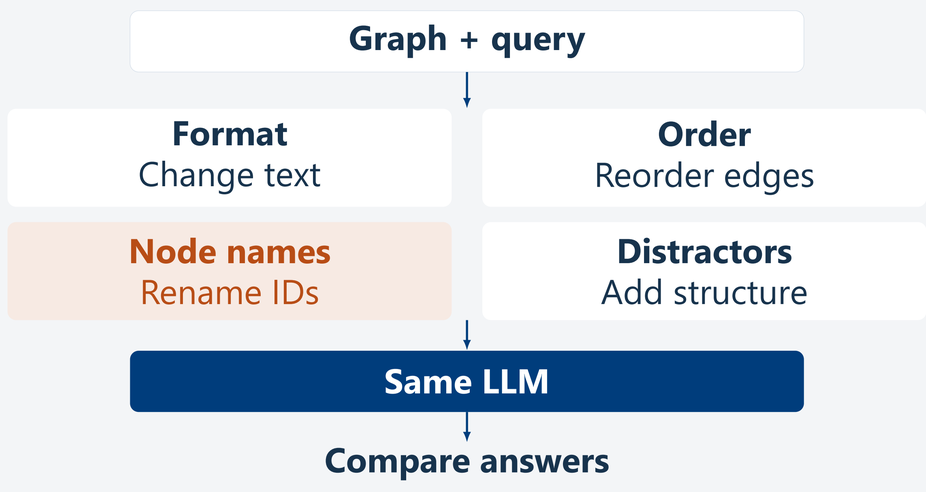
05 Benchmark

Task	Question	Label balance
Connectivity	Path exists?	300 yes / 0 no
Shortest path	How many edges?	Lengths 1–5: 101/143/50/4/2
Triangle	Node in a triangle?	184 yes / 116 no
Common neighbor	Shared neighbor?	202 yes / 98 no

100 connected ER graphs · 5–15 nodes · $p = .3$ · seed 42
1,200 queries · temperature 0 · yes/no or integer parser

Names: generic IDs vs. degree-sorted semantic names. Graph and query IDs are remapped together.

06 Audit protocol



07 Future work

Larger graphs
Test larger, real-world graphs and richer query types.

Controlled names
Compare arbitrary-ID instructions with role-congruent and conflicting names.

Improve invariance
Test canonicalization and training with relabeled graphs.