

Graph Evidence Is Not Text Evidence: Serialization Faithfulness as Structural Invariance in Graph-Grounded Language Models

Lifan Hu

Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore 117417
lifan.hu@u.nus.edu

Abstract

Graph-grounded language models are usually evaluated by answer accuracy, yet a serialized prompt is a string representation of a graph rather than the graph itself. We argue that faithful graph-grounded reasoning requires the composed predictor $F_\theta \circ z$ to be well-defined on graph-query equivalence classes: structurally equivalent graph-query pairs should induce the same prediction regardless of textual realization. We formalize this as a structural invariance property, decompose unfaithfulness into syntactic instability, semantic interference, and distractor susceptibility, and introduce corresponding diagnostics. On controlled graph reasoning tasks, DeepSeek-V4-Flash achieves 98.9% accuracy and near-perfect syntactic stability ($FS = .975$), yet its isomorphism consistency drops from .969 to .701 under semantic-name relabeling ($\Delta_{sem} = .268$); only 72.9% of originally correct answers survive this transformation. Across four model families, we observe targeted semantic interference in high-accuracy models, broad unfaithfulness in weaker models, and near-perfect invariance as a capability ceiling. A graph-grounded LLM can be accurate without being structurally faithful: its predictions may fail to remain invariant under graph-equivalent textual relabelings.

1 Introduction

Graph-grounded language models are used for graph reasoning, knowledge graph QA, and retrieval-augmented generation (Edge et al., 2024; He et al., 2024b; Tang et al., 2024). But these models consume text, not graphs; any graph evidence must pass through a *textual serialization* step, typically treated as a neutral implementation detail. We argue this is unwarranted. Citation faithfulness research establishes that an answer can be correct but not faithfully grounded in the provided evidence (Wallat et al., 2025); the same distinction

applies to graph evidence. An answer produced by relying on node names rather than connectivity may be correct, but it is not structurally faithful.

This paper makes three contributions. First, we formalize serialization faithfulness as an invariance property over graph-query equivalence classes, establishing structural invariance as a necessary condition for graph faithfulness (Proposition 1). Second, we decompose unfaithfulness into syntactic instability, semantic interference, and distractor susceptibility. Third, we show that accuracy does not imply structural faithfulness: a model can answer correctly yet flip its answer under isomorphic relabeling with semantically suggestive names.

2 Related Work

Graph-grounded LLMs. GraphGPT (Tang et al., 2024) aligns GNN-encoded tokens with LLM tokens; InstructGLM (Ye et al., 2024) encodes structure via language; GraCoRe (Yuan et al., 2025) notes “node ordering impacts task success”; TalkLikeaGraph (Fatemi et al., 2024) finds format-dependent accuracy variation; NLGraph (Wang et al., 2023) shows brittleness to spurious correlations. These establish that representation matters, but primarily evaluate accuracy. We isolate graph-preserving transformations as diagnostics for faithful interpretation.

Faithfulness and sensitivity. Citation correctness $\neq$ faithfulness (Wallat et al., 2025); input ordering can substantially affect LLM outputs (Guan et al., 2025); format sensitivity is a general LLM property (He et al., 2024a); in graph ML, permutation-invariant or permutation-equivariant computation is a fundamental requirement (Zhang et al., 2023). We connect these at the graph-to-language interface: does serialized prompting induce a well-defined graph reasoning function?

3 Serialization Faithfulness as Graph Invariance

A graph reasoning task defines a target function $y : \mathcal{G} \times \mathcal{Q} \rightarrow \mathcal{Y}$ over graph-query pairs. A serialization-based LLM instead computes a function over strings, $F_\theta : \Sigma^* \rightarrow \mathcal{Y}$. Serialization faithfulness asks whether $F_\theta \circ z$ is well-defined on graph-equivalence classes: equivalent graph-query pairs should induce the same prediction regardless of their textual realization.

Formal setup. A serialization map $z_{\rho,\pi,\ell} : \mathcal{G} \times \mathcal{Q} \rightarrow \Sigma^*$ depends on format ρ , ordering π , and labeling ℓ . We suppress the serialization subscripts when they are clear from context. *Structural equivalence*: $(G, q) \sim_s (G', q')$ if $G' = \phi(G)$ is an isomorphic relabeling and $q' = \phi(q)$. *Label-preserving equivalence*: $(G, q) \sim_\ell (G', q')$ if $y(G, q) = y(G', q')$. IC uses $\sim_s$; DR uses $\sim_\ell$. This distinction separates whether a model respects structural equivalence from whether it is robust to weaker answer-preserving perturbations. A model is serialization-faithful with respect to an equivalence $\sim$ if

$$(G, q) \sim (G', q') \Rightarrow F_\theta(z(G, q)) = F_\theta(z(G', q')) \quad (1)$$

Desideratum: well-defined graph reasoning under serialization. For a serialized LLM to behave as a graph-grounded reasoner, the composed predictor $F_\theta \circ z$ should be well-defined on graph-query equivalence classes. That is, if two serialized strings encode structurally equivalent graph-query pairs, then any difference in prediction reflects dependence on non-structural features of the serialization rather than on the graph evidence itself.

Proposition 1 (Structural invariance is necessary). Let h be a predictor whose value depends only on the isomorphism class of a graph-query pair (G, q) . Then for any isomorphism ϕ , $h(G, q) = h(\phi(G), \phi(q))$. Consequently, if $F_\theta \circ z$ changes its prediction under an isomorphic relabeling, the change cannot be explained by graph structure alone. *Proof.* An isomorphism preserves all structural relations (adjacency, paths, degrees, motifs). Since h depends only on the isomorphism class, relabeling identifiers cannot change its output; any prediction change must arise from non-structural features of the serialization.

Unfaithfulness decomposition. We decompose unfaithfulness into three components:

1. *Syntactic serialization instability* ($1 - \text{FS}$, OS): dependence on format or ordering.
2. *Semantic interference* ($\Delta_{\text{sem}} = \text{IC}_N - \text{IC}_M$): the gap between neutral and semantic-name relabeling.
3. *Distractor susceptibility* ($1 - \text{DR}$): sensitivity to label-preserving irrelevant structure.

The central component is semantic interference, measured by the consistency gap between neutral and semantic-name relabeling. Each diagnostic instantiates the same invariance requirement with a different transformation family: format stability varies the textual map z_ρ while holding the graph-query pair fixed; order sensitivity varies the ordering π ; isomorphism consistency varies the graph-query pair within structural equivalence $\sim_s$; distractor robustness varies the graph-query pair within label-preserving equivalence $\sim_\ell$. Thus all metrics test whether $F_\theta \circ z$ is well-defined under transformations that preserve the task-relevant graph evidence.

Isomorphism Consistency (IC). Whether structurally identical graphs with relabeled nodes yield the same answer: $\text{IC}(q, G, \phi) = 1[\hat{a}(q, z(G)) = \hat{a}(\phi(q), z(\phi(G)))]$. We test *neutral* relabeling ($A \rightarrow X_1, B \rightarrow X_2, \dots$) yielding IC_N , and *semantic-name* relabeling yielding IC_M , which tests whether semantically suggestive identifiers affect predictions on an isomorphic graph. We define:

$$\Delta_{\text{sem}} = \text{IC}_N - \text{IC}_M \quad (\text{semantic interference}) \quad (2)$$

$$\text{CF}_M = \Pr[\hat{a}_M = y \mid \hat{a}_0 = y] \quad (3)$$

CF_M measures whether predictions that were initially correct remain correct after semantic relabeling, a stricter test than consistency alone. $\text{CF}_M < 1$ indicates dependence on surface cues disrupted by the relabeling.

Supporting diagnostics. *Format Stability* (FS): agreement across formats; *Order Sensitivity* (OS): answer flip rate under reordering; *Distractor Robustness* (DR): whether label-preserving disconnected structure changes the prediction ($y(G, q) = y(G \cup D, q)$ required).

4 Experimental Setup

Tasks and graphs. Four graph reasoning tasks with structurally defined ground truth, on 100 con-

Factor	Settings
Tasks	Connectivity, Path, Triangle, Common Neigh.
Graphs	100 ER graphs, $n \in \{5, \dots, 15\}$, $p=0.3$
Models	Qwen3.6-35B-A3B, DeepSeek-V4-Flash, Llama-4-Scout, gpt-5.4-mini, gpt-5.4
Formats	Edge list, Adj. list, JSON, NL
Relabeling	Neutral ($X_1, X_2, \dots$), semantic-name
Distractors	Disconn. comp., $\triangle$, high-deg. star

Table 1: Experimental conditions. All distractors are verified to be label-preserving.

nected Erdős–Rényi graphs ($n \in \{5, \dots, 15\}$, $p = 0.3$), seeded deterministically, yielding 400 graph-task pairs and three queries per pair (1,200 query instances). The observed target distribution is 300 yes and 0 no for connectivity; shortest-path distances 1–5 occur 101/143/50/4/2 times; triangle labels contain 184 yes and 116 no; common-neighbor labels contain 202 yes and 98 no.

Models. We evaluate three open-weight models from distinct families: Qwen3.6-35B-A3B (Alibaba; 35B total/3B active; Apache-2.0), DeepSeek-V4-Flash (DeepSeek; 284B/13B active; MIT), and Llama-4-Scout (Meta; 109B/17B active; Llama license). We additionally evaluate the proprietary gpt-5.4-mini-2026-03-17 and use gpt-5.4-2026-03-05 as a capability ceiling. All models use temperature 0. The Qwen and DeepSeek API configurations explicitly disable reasoning; no explicit reasoning setting is passed for gpt-5.4. Outputs are mapped by a deterministic task-specific parser to yes/no or an integer; unmatched outputs receive the sentinel unclear. Evaluation code is available at https://github.com/Anormalm/graph_evidence.

Relabeling. Neutral relabeling maps all sorted nodes to generic identifiers ($X_1, X_2, \dots$). Semantic-name relabeling automatically sorts all nodes by decreasing degree and assigns, in order, Hub, Center, Bridge, Anchor, Core, Leaf, Edge, Tip, Outlier, Tail, then Node11–Node15. The synthetic graphs contain no semantic node attributes; the names appear only as textual identifiers. Canonical edge order is recomputed after relabeling, so IC includes any induced lexical-order change.

Aggregation. FS averages all six unordered pairs of the four formats. OS uses canonical, one seeded-random, and one BFS ordering; the archived implementation compares all three with canonical,

Model	Acc.	1–FS↓	OS↓	$\Delta_{\text{sem}}\downarrow$	1–DR↓	$\text{CF}_M\uparrow$
Symbolic oracle	1.00	.000	.000	.000	.000	1.00
gpt-5.4-mini	.936	.084	.046	.152	.066	.533
Qwen3.6-35B-A3B	.541	.325	.086	.061	.318	.575
DeepSeek-V4-Flash	.989	.025	.018	.268	.026	.729
Llama-4-Scout	.689	.259	.142	.091	.306	.658

Table 2: Main results. Columns match the unfaithfulness decomposition: syntactic instability (1–FS, OS), semantic interference (Δ_{sem}), distractor susceptibility (1–DR), and robust correctness (CF_M). The oracle computes each task from adjacency, ignoring node-name semantics.

including its zero self-comparison. IC compares one neutral and one semantic-name relabeling with the original, and DR compares three distractor conditions with the baseline. Reported aggregates pool comparison indicators across the four equally sized tasks rather than macro-averaging task means. The implementation computes 95% percentile intervals from 1,000 seed-42 bootstrap resamples of comparison indicators; Table 2 reports point estimates. The shared prompt asks the model to reason from the graph description but does not explicitly state that identifiers are arbitrary.

5 Results

Table 2 presents the main results.

Accuracy does not imply faithfulness. gpt-5.4-mini achieves 93.6% accuracy, yet $\text{CF}_M = .533$: only 53.3% of correct answers survive semantic-name relabeling. DeepSeek-V4-Flash reaches 98.9% accuracy but still loses 27.1% of correct answers ($\text{CF}_M = .729$), with $\Delta_{\text{sem}} = .268$, the highest semantic interference observed. When semantic identifiers are introduced, we observe prediction changes even though the underlying graph is preserved.

Three failure profiles. *Targeted semantic interference:* DeepSeek-V4-Flash has excellent syntax (1–FS = .025, OS = .018) and distractor robustness (1–DR = .026), but degrades sharply under semantic-name relabeling ($\text{IC}_N = .969 \rightarrow \text{IC}_M = .701$, $\Delta_{\text{sem}} = .268$). gpt-5.4-mini shows a similar but milder pattern ($\Delta_{\text{sem}} = .152$). *Broad unfaithfulness:* Qwen3.6-35B-A3B and Llama-4-Scout show degradation across all diagnostics (1–FS = .325 / .259, 1–DR = .318 / .306), with

small Δ_{sem} (.061 / .091) only because they are already unfaithful under neutral relabeling; their CF_M (.575 / .658) is partly sustained by the subset of easy tasks (connectivity) where relabeling matters less. For these lower-accuracy models, small Δ_{sem} should not be read as semantic robustness: when neutral consistency is already weak, the semantic gap is compressed by broad instability. The dominant failure mode shifts with capability: weaker models fail broadly, while stronger models fail specifically at semantic interference, and this failure is invisible to accuracy alone.

Capability ceiling. gpt-5.4 achieves near-perfect invariance ($\Delta_{\text{sem}} = 0$, $CF_M = 1.0$, $FS = .989$), suggesting that serialization faithfulness may improve with model capability. Notably, $CF_M < 1$ is possible even at $> 93\%$ accuracy, undetectable by accuracy alone.

6 Conclusion

We have shown that accuracy does not imply structural faithfulness. Proposition 1 establishes structural invariance as necessary; we decomposed unfaithfulness into syntactic instability, semantic interference, and distractor susceptibility. Across four model families, we observe three failure profiles: targeted semantic interference in high-accuracy models (DeepSeek: $\Delta_{\text{sem}} = .268$, $CF_M = .729$ at 98.9% accuracy; gpt-5.4-mini: $\Delta_{\text{sem}} = .152$, $CF_M = .533$), broad unfaithfulness in weaker models (Qwen3.6, Llama-4-Scout), and near-perfect invariance as capability ceiling (gpt-5.4: $\Delta_{\text{sem}} = 0$). A graph-grounded LLM can be accurate without being structurally faithful.

Limitations

This study has several limitations. First, we evaluate on synthetic graphs of 5–15 nodes; effects may differ at scale. Second, our four formats do not cover all representations. Third, we extract answers via pattern matching. Fourth, we diagnose but do not propose improvement methods. Fifth, two models are proprietary; we include three open-weight models for broader model-family coverage. Sixth, the prompt does not explicitly declare identifiers arbitrary, and we do not include a separately designed role-congruent control; semantic-name sensitivity therefore does not by itself separate cue integration from instruction-following. Seventh, all generated graphs are connected, making the connectivity subset label-degenerate, and we do not report chance-

corrected agreement. Eighth, the archived metric implementation handles unclear outputs differently across diagnostics, and two named distractor conditions instantiate the same disconnected triangle; these implementation details limit fine-grained interpretation of the point estimates.

Ethical Considerations

This work uses synthetic graphs and no human or personal data. Its primary foreseeable risk is over-generalization: results from controlled prompts are not safety or reliability guarantees for deployed systems.

Acknowledgments

This research was motivated by a practical problem encountered during the author’s industry internship about graph representation and fraud detection.

References

- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. [From local to global: A Graph RAG approach to query-focused summarization](#). *Preprint*, arXiv:2404.16130.
- Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. [Talk like a graph: Encoding graphs for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. 2025. [The order effect: Investigating prompt sensitivity to input order in LLMs](#). *Preprint*, arXiv:2502.04134.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024a. [Does prompt formatting have any impact on LLM performance?](#) *Preprint*, arXiv:2411.10541.
- Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024b. [G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering](#). In *Advances in Neural Information Processing Systems 37*, pages 132876–132907. Neural Information Processing Systems Foundation, Inc.
- Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. [GraphGPT: Graph instruction tuning for large language models](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, pages 491–500, New York, NY, USA. Association for Computing Machinery.

- Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. [Correctness is not faithfulness in retrieval augmented generation attributions](#). In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, pages 22–32, New York, NY, USA. Association for Computing Machinery.
- Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. [Can language models solve graph problems in natural language?](#) In *Advances in Neural Information Processing Systems 36*, pages 30840–30861. Neural Information Processing Systems Foundation, Inc.
- Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2024. [Language is all a graph needs](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1955–1973, St. Julian's, Malta. Association for Computational Linguistics.
- Zike Yuan, Ming Liu, Hui Wang, and Bing Qin. 2025. [GraCoRe: Benchmarking graph comprehension and complex reasoning in large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7925–7948, Abu Dhabi, UAE. Association for Computational Linguistics.
- Bohang Zhang, Guhao Feng, Yiheng Du, Di He, and Liwei Wang. 2023. [A complete expressiveness hierarchy for subgraph GNNs via subgraph Weisfeiler–Lehman tests](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41019–41077. PMLR.