Do LLMs Follow Their Self-Reported Causal Graphs? A Graph-Contract Audit of Falsifiable Rationales for Trustworthy Decisions

Anonymous Authors[1]

Abstract

LLMs are increasingly considered for decision-support settings where explanations must support accountability, not merely plausibility. Causal graphs are a natural interface for such explanations: they state which variables should matter, which should not, and which information should be inadmissible. But do LLMs actually follow the causal graphs they report? We propose a graph-contract audit that treats a self-reported causal graph as a falsifiable behavioral commitment. An LLM first reports a task-level causal graph; later, in fresh prediction prompts without access to that graph, we test whether its decisions respect the graph-implied constraints. Our audit checks whether predictions remain stable under graph-implied irrelevant perturbations and whether models avoid using post-outcome information. Across admissions and loan-default decision tasks, we find that self-reported causal graphs often generate meaningful, auditable commitments, but model predictions frequently violate them. These violations are not fully explained by predictive accuracy, graph stability, decoding noise, or explicit graph reminders. Our results suggest that causal explanations for LLM decisions should be evaluated as testable contracts. This provides a practical framework for turning causal explanations into accountability tools for trustworthy AI in socially consequential decision workflows.

1. Introduction

Large language models (LLMs) are increasingly considered for decision-support workflows where an explanation must do more than sound plausible. In admissions, lending, benefit allocation, or similar oversight settings, a decision aid may be asked to state which variables should matter, which variables should not matter, and which information should be unavailable at decision time. A causal graph is an appealing interface for this purpose: unlike a free-form rationale or flat feature list, it makes structured commitments about conditional irrelevance, mediated influence, and post-outcome inadmissibility. The question is therefore concrete: *do LLMs follow the causal graphs they report?*

This question differs from asking whether an LLM knows causal facts or can reason with a graph supplied in the prompt. A self-reported causal graph is a public rationale. It may be useful for accountability even when it is not the true causal graph of the world, and it may be misleading even when it looks substantively reasonable. The relevant audit target is behavioral: if the model reports a graph that says a field should be irrelevant, or that post-outcome information should not affect an ex ante score, then later predictions should respect that commitment when the graph is no longer shown.

We introduce a *graph-contract audit* for this setting. The model first reports a task-level directed acyclic graph (DAG) over a fixed variable schema. That conversation is discarded. The evaluator parses the graph, derives graph-implied commitments, builds matched profile pairs that differ only in the audited field, and queries the model in a fresh graph-absent prediction context. The graph is therefore used only to construct the audit contract, not as an input to the primary prediction prompt. We measure noise-corrected score shifts and decision flips, while reporting coverage as a guardrail so that untestable graphs do not look faithful by default.

Across admissions and loan-default tasks, the audit finds a persistent mismatch between reported causal rationales and later behavior. In a 192-cell balanced core grid, 48 of 96 admissions cells and 70 of 96 loan-default cells license at least one graph-implied test. Violations persist after repeat-noise correction. Role-conditioned estimates show that ordinary irrelevance, demographic-field irrelevance, and post-outcome leakage tests behave differently. Graph reminders and replayed context move scores, but do not certify that the earlier graph faithfully explains fresh predictions.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

This makes the work a natural fit for trustworthy AI evaluation in socially consequential decision support. The method does not certify deployment, legal compliance, or population causal effects. It provides a narrower but useful standard: causal rationales should be testable, and contradictions between a model's public graph and its later decisions should be visible before that graph is used for accountability.

Contributions. First, we define fresh-conversation graph-contract faithfulness: a self-reported causal graph is treated as a falsifiable public rationale, not as ground truth or an internal mechanism. Second, we turn graph semantics into matched behavioral tests for irrelevant-field omission, irrelevant-value swaps, and post-outcome leakage, with coverage and repeat-noise guardrails. Third, we instantiate the audit on two decision tasks and show that accuracy, graph stability, explicit graph access, and scale are not substitutes for behavioral faithfulness tests.

2. Related Work and Positioning

LLMs and causal graphs. Recent work evaluates LLMs as causal reasoners, causal-graph users, or assistants for causal discovery and causal inference (Kiciman et al., 2023; Gao et al., 2023; Jin et al., 2023; 2024; Chen et al., 2024; Zhou et al., 2024; Zecevic et al., 2023; Sheth et al., 2024; Wan et al., 2025; Liu et al., 2025). These studies ask whether models can answer causal questions, infer or use causal structure, or exploit graphs that are supplied as task inputs. Our question is different. We ask whether a graph emitted by the model as its own task rationale remains behaviorally binding when the model later predicts without seeing that graph.

This distinction matters for accountability. A model can use a graph well when it is printed in the prompt, yet fail to behave consistently with a graph it previously offered as an explanation. Conversely, a reported graph can be substantively imperfect and still license useful tests of whether the model honors its own stated exclusions. We therefore avoid treating graph quality, graph use, and graph-faithfulness as interchangeable quantities.

Faithfulness of explanations. Explanation faithfulness is the relation between an explanation and the behavior it purports to explain, not the explanation's surface plausibility (Jacovi & Goldberg, 2020). LLM rationales, including chain-of-thought traces, can be fluent while failing to causally support the answer (Lanham et al., 2023; Turpin et al., 2023; Paul et al., 2024; Matton et al., 2025; Xiong et al., 2025; Shen et al., 2025; Somov et al., 2026). Causal graphs sharpen this problem because they have semantics: they imply testable claims about what should and should not affect a score. Our audit uses those semantics rather than treating the graph as another piece of persuasive text.

Behavioral audits. Behavioral testing frameworks such as CheckList convert expected invariances into tests (Ribeiro et al., 2020). In our setting the expected invariances are endogenous to the model's own graph report. This makes the test useful for oversight: the evaluator is not merely asking whether a model passes externally authored invariance tests, but whether it honors commitments implied by the rationale it chose to present.

The workshop setting also motivates a narrower standard than deployment approval. A public agency, lender, or admissions office may not need to know a model's internal circuit to reject a bad explanation interface. It is enough to show that the interface makes commitments that later behavior contradicts. Graph-contract auditing supplies exactly that intermediate evidence standard.

3. Graph-Contract Audit

The audit starts from a simple contract view. When a model reports a causal graph for a decision task, the graph is not treated as a discovered world model or as a readout of hidden model state. It is treated as a public rationale. Public rationales can be checked: if the graph says a field is irrelevant after conditioning on the rest of the profile, then changing only that field should not move the later score beyond the model's own repeat noise. If the graph places a variable after the decision outcome, then revealing that variable should not alter an ex ante score.

Figure 1 summarizes the procedure. The fixed task schema separates pre-decision variables, post-outcome variables, and the binary target Y. A graph report is valid if it is parseable, uses only allowed variables, and is acyclic. Invalid reports are counted as explanation-interface failures rather than silently discarded. The evaluator uses valid graphs to build tests, but the primary prediction prompt never shows the graph. This separation is what makes the test a faithfulness audit rather than a graph-following prompt. Table 1 gives a concrete prompt trace so that the separation between graph reporting, fresh prediction, and graph-access controls is visible in text rather than only in notation.

For a pre-decision field A, the graph licenses an irrelevance test when Y is d-separated from A given the other pre-decision fields in the reported graph. The evaluator then compares matched profiles (x, x') that differ only in A, either by omitting A or swapping its value on support. For a post-outcome variable Z, the graph licenses a leakage test when Z is represented as downstream of the target but the prediction task is explicitly ex ante. In that case, revealing Z should not alter the fresh prediction score.

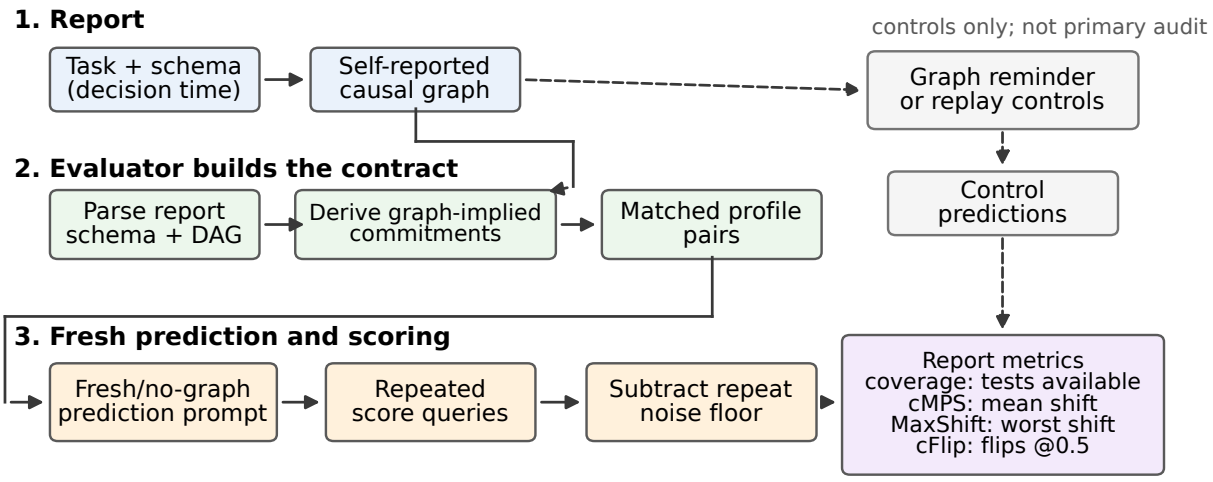

Figure 1. Graph-contract audit. The LLM first reports a causal graph. The evaluator parses it, derives graph-implied commitments, and constructs matched perturbation pairs. The primary prediction prompt is fresh and graph-absent; graph reminders and replayed context are controls. Coverage is the fraction of candidate commitments that become testable. cMPS is noise-corrected mean probability shift, MaxShift is the largest corrected test-level shift, and cFlip is the corrected decision-flip rate at threshold 0.5.

Table 2 is the bridge between graph semantics and behavioral tests. The omission and swap tests are ordinary conditional-invariance checks induced by the reported graph. The post-outcome test is different: it is tied to the decision time. A field observed only after the decision is inadmissible for the ex ante score even if the model can use it when it appears in the prompt.

Let $S_M(x; C, T_p) \in [0, 1]$ be the score returned by model M for profile x, prediction context C, and prediction temperature T_p. For a licensed test c, the raw mean probability shift is

$$\Delta_c = \mathbb{E}|S_M(X; C, T_p) - S_M(X'; C, T_p)|.$$

Repeated same-input score queries estimate a noise floor ν_c. We report the excess shift $\Delta_c^+ = [\Delta_c - \nu_c]_+$. The main behavioral error is $\mathrm{cMPS} = N^{-1} \sum_c \Delta_c^+$ over licensed tests. MaxShift is the largest Δ_c^+, and cFlip is the analogous noise-corrected decision-flip rate at threshold 0.5. Coverage is reported separately because a graph with no licensed tests can have zero shift without being informative.

Two design choices keep the estimand narrow. First, the matched-pair tests hold the rest of the structured profile fixed, so a score movement is evidence against the reported graph contract for that field, not an estimate of a population intervention effect. Second, invalid graph reports are not folded into the denominator of a behavioral shift estimate. They are a separate failure mode of the explanation interface: the model did not produce a valid directed acyclic graph over the declared schema.

The audit therefore has a clear null interpretation. Under an ideal scorer that follows the reported graph semantics, licensed irrelevance and post-outcome tests have zero excess shift after repeat-noise correction. A nonzero excess shift is a behavioral contradiction of the report. A zero shift with no licensed tests is different: it says the reported graph did not expose that part of the model to a falsifiable check.

4. Experimental Design

We use two structured binary decision tasks: admissions and loan default. Both have fixed schemas and explicit decision times; only the admissions task has a known structural causal model used for a diagnostic graph-correctness check. The loan-default task is included because it resembles a common decision-support setting where post-outcome information, such as later missed payments, must not influence an underwriting-time score.

Table 3 gives the experimental vocabulary used throughout the paper. The main evidence comes from a balanced core grid. A grid cell fixes the task, model, graph-elicitation temperature T_g, and prediction temperature T_p. Within a cell, the model reports multiple graphs, the evaluator derives graph-implied tests, and predictions are queried repeatedly on matched profile pairs. The repeated score queries estimate decoding noise; they are not treated as independent model mechanisms. The primary protocol is fresh/no graph. Control protocols add the reported graph as a reminder or replay the graph-elicitation context before prediction.

The two tasks are intentionally similar in format but differ-

Table 1. Abridged prompt trace for one audit. The graph-report prompt elicits a task-level causal graph; the later prediction prompts are fresh and do not contain that graph. The displayed scores are one illustrative matched pair, not an aggregate result. In this pair the primary fresh/no-graph scores do not move, while the graph-reminder control does; aggregate violations are reported in Section 5.

Stage	Prompt or output excerpt	Audit role
Graph report prompt	You are advising on a loan-default risk score at underwriting time. The decision is made before later repayment outcomes are observed. Use only the listed fields and the target `default`. Report a directed acyclic causal graph over these variables.	Elicits a public causal rationale over the fixed schema and decision time.
Returned graph	Example arrows include `credit_limit` → `bill_ratio`, `repayment_status` → `default`, `bill_ratio` → `default`, and `payment_ratio` → `default`.	Parsed into graph-implied commitments such as field irrelevance or post-outcome inadmissibility.
Fresh prediction prompt	New conversation. Profile only; no graph text. The model is asked for a probability in $[0, 1]$ that the applicant defaults.	Primary behavioral audit. The graph is absent, so score shifts test consistency with the earlier rationale.
Matched perturbation	Same profile pair, but the audited field is omitted or swapped while all other fields remain fixed. Fresh/no-graph score: $0.257 \to 0.257$. Graph-reminder control: $0.789 \to 0.875$.	The primary pair has no score movement here; the graph-reminder movement illustrates why graph-access controls are analyzed separately.

Table 2. Audit families derived from a reported graph. Each row defines a licensed test family and its graph-contract null. The experimental unit is a matched profile pair generated from a valid graph report; zero excess shift is the ideal graph-contract behavior after repeat-noise correction.

Test family	When licensed	Expected behavior
Omission	$Y \perp_G A \mid B_A$ and the field can be omitted	Adding or removing A should not change the score beyond repeat noise.
Value swap	$Y \perp_G A \mid B_A$ and an on-support replacement exists	Changing only A's value should not change the score beyond repeat noise.
Post-outcome	The graph places Z downstream of Y, and Z is unavailable at decision time	A post-outcome variable should not alter an ex ante score.

ent in what can be checked. Admissions provides a controlled setting with a known schema and a reference structural causal model for descriptive graph-quality diagnostics. Loan default provides a decision-support setting closer to the accountability motivation: several fields are ordinary pre-decision fields, demographic fields require special care, and later repayment information is unavailable at underwriting time. This lets the same audit test ordinary irrelevance, demographic irrelevance, and post-outcome inadmissibility without changing the basic protocol.

The model panel is used as an evaluation panel, not as a ranking benchmark. It contains multiple open-weight model families and sizes so that the paper can ask whether graph-contract behavior is explained away by model scale, family, or decoding temperature. The answer is mixed enough that the paper avoids a scaling-law claim. Some rows are high-coverage and low-shift; others have high coverage and large MaxShift. This variation is useful: it shows that the audit is sensitive to behavior rather than merely reproducing a predetermined failure label.

The control protocols have a separate purpose from the primary audit. Fresh/no graph asks whether a graph emitted earlier constrains later predictions when the graph is absent. Fresh+graph reminder asks whether explicitly showing the graph changes the model's score. Replay/no extra graph asks whether reconstructing the previous graph-report context matters even without a new reminder. Replay+graph reminder combines both interventions. These controls are important for interpretation, but they are not the primary faithfulness target: a model can become more graph-consistent after being reminded of the graph while still failing the stricter fresh-conversation audit.

Table 4 is the first sanity check: the audit is often testable, but not universally so. Admissions has 48 testable cells out of 96, while loan default has 70 out of 96. The invalid-report counts, 8 and 12 cells respectively, are not hidden; they indicate that some graph reports fail the parser, schema, or acyclicity requirements. We therefore interpret low cMPS only together with coverage. Zero or tiny coverage is an availability diagnostic, not evidence that the model is graph-faithful.

5. Results

Auditable rationales and violations coexist. Figure 2A plots coverage against cMPS for the balanced grid. It contains both high-coverage low-shift rows and high-coverage higher-shift rows. This matters: the audit is not designed to force failure, and coverage by itself is not a certificate. The strongest reading is joint: a graph must license testable commitments, and fresh predictions must then respect those commitments after repeat-noise correction.

Table 3. Design and metric map. This table defines the main experimental units and avoids conflating faithfulness errors with guardrails or diagnostics. A licensed test is a graph-implied commitment that can be evaluated on matched profile pairs.

Quantity	Value or definition	Interpretation
Core grid	12 model variants $\times$ 2 tasks $\times$ 4 graph temperatures $\times$ 2 prediction temperatures $=$ 192 cells	Balanced main panel; each task contributes 96 cells.
Graph temperature T_g	$\{0.0, 0.2, 0.5, 0.8\}$	Controls diversity of self-reported graph rationales.
Prediction temperature T_p	$\{0.0, 0.5\}$	Controls stochasticity of later score predictions.
Graph reports	16 target reports per grid cell	Unit used to derive graph-implied commitments and invalid-report rates.
Repeated score queries	3 per prompt condition	Estimates repeat-noise floor for score shifts and decision flips.
Matched profile pairs	128 target pairs per licensed test	Profiles differ only in the audited field.
Coverage	Fraction of candidate commitments that become testable	Guardrail; higher means more of the reported graph can be audited.
cMPS / MaxShift / cFlip	Noise-corrected mean shift, maximum shift, and flip rate	Behavioral graph-contract errors; lower is more consistent.
AUROC	Area under the receiver operating characteristic curve	Accuracy diagnostic, not a graph-faithfulness metric.

Table 4. Balanced core-grid summary. The grid has 192 evaluated cells, 96 per task. Each cell targets 16 graph reports, three repeated score queries per prompt condition, and 128 matched profile pairs per licensed test. Coverage is a test-availability guardrail; cMPS is noise-corrected mean probability shift over licensed tests; AUROC is area under the receiver operating characteristic curve and measures predictive discrimination, not graph faithfulness.

Task	Cells	Testable	Invalid	Coverage ↑	cMPS ↓	AUROC ↑
Admissions	96	48	8	0.000–0.714	0.001–0.031	0.615–0.813
Loan default	96	70	12	0.000–0.722	0.001–0.030	0.458–0.613

Figure 2B shows why graph roles should not be collapsed into a generic perturbation score. Admissions post-outcome leakage has the largest role-conditioned shift in the displayed estimates, while loan-default post-outcome leakage is small but available in only 5 of 96 cells. Ordinary and demographic irrelevance also show nonzero excess shifts. These roles have different meanings for oversight. An ordinary-field violation says the model used a field its graph marked irrelevant. A demographic-field violation raises a stronger accountability question because the graph explicitly licensed invariance to a sensitive attribute. A post-outcome leakage violation means the score moved when information unavailable at decision time was introduced. The conclusion is not that every model violates every graph. It is that different graph-implied commitments produce different behavioral tests, and each must be audited on its own terms.

Accuracy, stability, and scale are not certificates. The evidence separates behavioral graph faithfulness from adjacent proxies. Loan-default Qwen2.5 14B reaches AUROC 0.613 while a representative high-coverage row still has cMPS 0.018 and MaxShift 0.732. A loan-default Qwen 14B row has coverage 0.722 and cMPS 0.002, showing that high-coverage low-shift rows exist. The scorecard below reports the same separation at a broader level: predictive discrimination, graph stability, invalid-report rate, coverage, and behavioral shift move separately. This is the point of the graph-contract framing. Accuracy and plausible structure are useful diagnostics, but neither tells us whether later decisions honor the reported causal rationale. Scale diagnostics are treated the same way: larger or newer models can be interesting comparisons, but parameter count is not itself an explanation-faithfulness measure.

Table 5 puts the diagnostic axes side by side. The admissions task has better predictive discrimination on average than loan default, but both tasks show nonzero behavioral shifts. Graph stability is also not enough: stable reported graphs can still license commitments that fresh predictions violate. The scorecard also records low-shift cases, because the audit is meant to distinguish faithful-looking cells from contradicted commitments rather than label every graph report as failing.

Graph access changes scores but does not certify faithfulness. Figure 3 compares graph-access controls to the fresh graph-absent baseline. Positive ΔcMPS means higher graph-contract violation than fresh. Replay lowers average cMPS for admissions, but loan-default graph reminders and replay increase cMPS and move scores substantially. The score-movement panel is important because a control can change predictions even when it does not repair the graph contract. The pattern is mixed, so the careful claim is not that reminders never help. It is that explicit graph access and reconstructed context are controls for steerability and

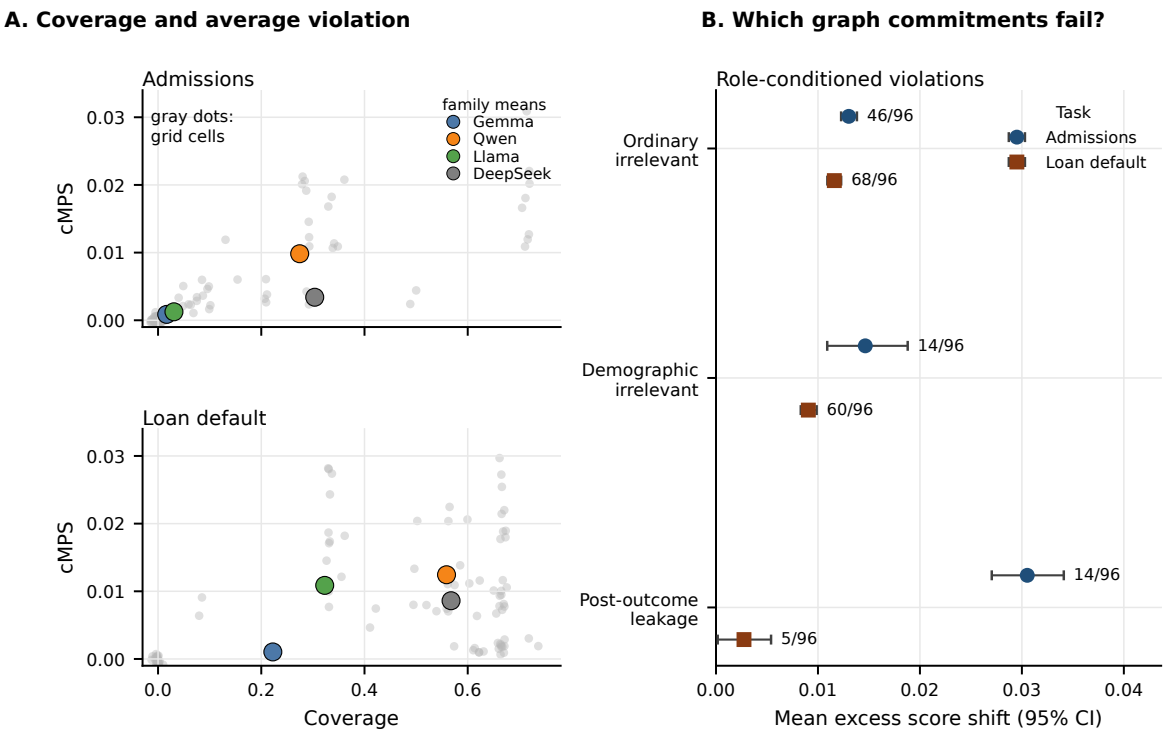

Figure 2. Coverage, role, and graph-contract violation. Panel A plots all 192 core-grid cells in gray and overlays task-specific model-family means to reduce visual clutter. Coverage is the fraction of candidate graph-implied commitments that become auditable; cMPS is lower-better noise-corrected mean probability shift over licensed tests. Panel B reports role-conditioned excess score shift with 95% bootstrap confidence intervals over graph-report units; labels give the number of grid cells with that licensed-test family out of 96 per task. The figure supports the main claim: self-reported graphs often create auditable commitments, and fresh predictions can violate those commitments in role-specific ways.

context dependence, not certificates that the graph was a faithful rationale for fresh predictions.

6. Interpreting the Audit

The audit is deliberately not a leaderboard. A model can look good under one diagnostic and poor under another because the diagnostics answer different questions. Coverage asks whether the reported graph exposed the model to falsifiable tests. cMPS, MaxShift, and cFlip ask whether fresh predictions obeyed those tests. AUROC asks whether the scores separate positive and negative labels. ECD asks whether repeated graph reports stabilize on similar test sets. The paper's main claim depends on keeping these questions separate.

This separation is most important for three common misreadings. First, a low shift is reassuring only when the graph licensed enough tests. A model that reports a graph with no usable irrelevance or leakage commitments can obtain a trivial zero shift, but that is an absence of audit evidence rather than a faithful rationale. Second, a high AUROC row is not a faithful-explanation row by default. Predictive discrimination says that the scores track labels; it does not say whether the score respects the model's stated exclusions. Third, graph reminders are not repairs unless they lower graph-contract errors without simply steering the model into a different scoring regime. Figure 3 shows why this distinction matters: controls can move scores substantially.

The role-conditioned results also clarify what is being falsified. An ordinary irrelevant-field test checks a graph-implied invariance for a non-sensitive field. A demographic-field test checks an invariance that is often more salient for oversight. A post-outcome leakage test checks whether information unavailable at decision time changes an ex ante score. These are not interchangeable failure modes. Combining them into a single perturbation number would hide whether the model contradicts a routine modeling exclusion, a demographic exclusion, or a time-inadmissibility commitment.

The practical reading is therefore contract-based. If an LLM presents a graph as a rationale, an auditor can ask: which parts of this graph are testable; which matched perturbations follow from those parts; and how much do fresh graph-absent scores move when those perturbations are applied? The answer may be favorable for some cells. That is use-

Table 5. Causal-consistency scorecard. Expected claim distance (ECD) summarizes graph-report stability; cMPS is noise-corrected mean probability shift; AUROC is predictive discrimination; CI means bootstrap confidence interval over graph-report units; range is a descriptive span over evaluated core-grid cells. The table has 96 cells per task and is a synthesis aid, not a ranking objective.

Quantity	Admissions	Loan default
Graph stability	ECD 0.040; range 0.000–0.214	ECD 0.057; range 0.000–0.318
Behavioral violation	cMPS 0.010; CI 0.008–0.012	cMPS 0.011; CI 0.009–0.013
Predictive accuracy	AUROC 0.716; range 0.615–0.813	AUROC 0.522; range 0.458–0.613
Ordinary irrelevant fields	shift 0.013; CI 0.012–0.014	shift 0.012; CI 0.011–0.012
Demographic irrelevant fields	shift 0.015; CI 0.011–0.019	shift 0.009; CI 0.008–0.010
Post-outcome leakage	shift 0.031; CI 0.027–0.034	shift 0.003; CI 0.000–0.005
Temperature sensitivity	cMPS 0.001–0.031; coverage 0.000–0.714	cMPS 0.001–0.030; coverage 0.000–0.722

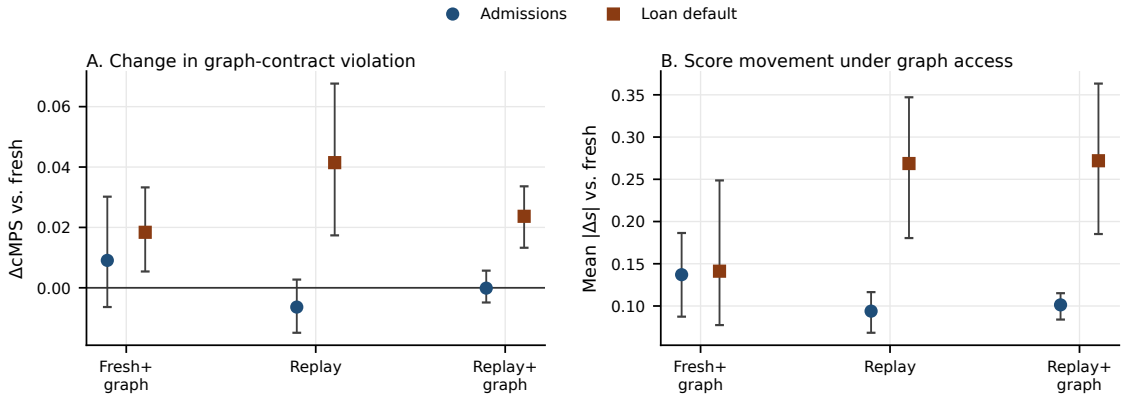

Figure 3. Graph-use controls relative to fresh graph-absent prediction. Panel A shows mean change in cMPS relative to fresh/no graph; positive values indicate higher graph-contract violation. Panel B shows mean absolute score movement $|\Delta s|$ on the same matched profiles. Intervals are 95% nonparametric bootstraps over six row-level units per task. The exact aggregate contains 2,000 admissions matched pairs and 4,400 loan-default matched pairs. Graph reminders and replay can move scores, but they do not certify behavioral faithfulness.

ful, because it means the audit can identify graph reports whose behavioral commitments are comparatively stable. The answer may also reveal contradictions. That is the accountability value: the failure is tied to a specific graph-implied commitment rather than to a vague sense that the explanation sounded unconvincing.

7. Limitations and Broader Impact

The audit assumes a fixed schema and explicit decision time. It does not handle open-ended dialogue where the relevant variables change during interaction. Value-swap tests on structured profiles are behavioral consistency tests, not population causal-effect estimates. The loan-default task is a realistic structured setting without a unique reference DAG, so the paper avoids claiming graph correctness there. More broadly, the reported graph is a public rationale, not a recovered internal mechanism.

The method is strongest when the task designer can say which variables are available before the decision and which variables arrive later. That is common in many administrative settings, but it is not universal. In more open-ended decision workflows, the first step would be a schema-design problem: decide what fields are in scope, what the decision time is, and which swaps remain on support. The graph-contract audit does not remove that burden. It makes the burden explicit so that a graph rationale can be checked against a declared task rather than judged only by plausibility.

The perturbation tests also have a local interpretation. When the score changes after an irrelevant-field swap, the audit has found a contradiction with the reported graph under the matched-profile construction. It has not estimated what would happen to a real population if that attribute changed, and it has not established a legal disparate-impact claim. This is an advantage for a workshop paper on trustworthy evaluation: the result is narrow enough to be actionable. An organization can ask for a causal rationale, derive the implied behavioral checks, and reject explanations that fail those checks without pretending that the audit solves deployment governance.

For AI for good applications, the positive use is accountability: a causal rationale can be converted into tests before it is relied on in a socially consequential workflow. The main risk is overinterpretation. Passing these audits is not a

deployment certificate, a legal fairness audit, or proof that a model has learned a human-acceptable causal mechanism. Failing an audit is narrower but important evidence: the model's later behavior contradicted a commitment implied by its own stated graph.

8. Conclusion

Self-reported causal graphs should not be treated as self-authenticating explanations. They are better viewed as falsifiable contracts. Our audit elicits a graph, removes it from primary prediction, derives graph-implied perturbation tests, and measures whether fresh predictions obey those tests after repeat-noise correction. Across two decision tasks, LLMs often produce graphs that are testable, yet later predictions can violate the resulting commitments. Trustworthy decision support therefore needs behavioral audits of causal rationales, not only plausible rationales, accurate predictions, or graph reminders.

References

Chen, S., Peng, B., Chen, M., Wang, R., Xu, M., Zeng, X., Zhao, R., Zhao, S., Qiao, Y., and Lu, C. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*, 2024.

Gao, J., Ding, X., Qin, B., and Liu, T. Is chatgpt a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11111–11126, 2023. doi: 10.18653/v1/2023.findings-emnlp.743.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. Cladder: Assessing causal reasoning in language models. *arXiv preprint arXiv:2312.04350*, 2023.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., and Schölkopf, B. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2024.

Kiciman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., Yu, T., McAuley, J., Ai, W., and Huang, F. Large language models and causal inference in collaboration: A survey. *arXiv preprint arXiv:2403.09606*, 2025.

Matton, K., Ness, R. O., Guttag, J., and Kiciman, E. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4ub9gpx9xw.

Paul, D., West, R., Bosselut, A., and Faltings, B. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024. doi: 10.18653/v1/2024.findings-emnlp.882.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Shen, X., Wang, S., Tan, Z., Yao, L., Zhao, X., Xu, K., Wang, X., and Chen, T. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2510.04040*, 2025.

Sheth, I., Fatemi, B., and Fritz, M. Causalgraph2llm: Evaluating LLMs for causal queries. *arXiv preprint arXiv:2410.15939*, 2024.

Somov, O., Chaichuk, M., Seleznyov, M., Panchenko, A., and Tutubalina, E. Breaking the chain: A causal analysis of LLM faithfulness to intermediate structures. *arXiv preprint arXiv:2603.16475*, 2026.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

Wan, G., Lu, Y., Wu, Y., Hu, M., and Li, S. Large language models for causal discovery: Current landscape and future directions. *arXiv preprint arXiv:2402.11068*, 2025.

Xiong, Z., Chen, S., Qi, Z., and Lakkaraju, H. Measuring the faithfulness of thinking drafts in large reasoning models. *arXiv preprint arXiv:2505.13774*, 2025.

Zecevic, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 2023.

Zhou, Y., Wu, X., Huang, B., Wu, J., Feng, L., and Tan, K. C. Causalbench: A comprehensive benchmark for causal learning capability of LLMs. *arXiv preprint arXiv:2404.06349*, 2024.

A. Formal Audit Properties

The propositions below describe the reported-semantics contract induced by a graph. They do not assert that the model internally computes with this graph.

Assumption 1 (Reported-semantics scorer)**.** Fix a valid reported graph G over the task schema. Let P_G be any distribution Markov to G. The ideal reported-semantics ex ante score is $r_G(x_{\mathrm{pre}}) = P_G(Y = 1 \mid x_{\mathrm{pre}})$, evaluated only on variables available at decision time.

Proposition 1 (Graph-implied null effects)**.** *If G licenses an omission or value-swap test for pre-decision variable A because $Y \perp_G A \mid B_A$, then $r_G(x_{B_A}, a) = r_G(x_{B_A})$ for admissible a. If G licenses a post-outcome leakage test for Z, then $r_G(x_{\mathrm{pre}}, z) = r_G(x_{\mathrm{pre}})$.*

Proposition 2 (Why coverage is necessary)**.** *If a graph licenses no tests, the audit reports no behavioral shift. This is a no-test condition, not evidence that the model's decisions are faithful to the graph.*

B. Exact Values Behind the Main Figures

Table 6 gives representative exact values behind the coverage panel. The figure itself plots the full 192-cell grid.

C. Prompt and Parser Summary

The graph-elicitation prompt gives the model the task, decision time, target, and fixed variable schema, then asks for a causal graph over exactly those variables. The parser accepts only reports that use allowed variable names and form an acyclic graph. The fresh prediction prompt presents a profile and asks for a probability score in $[0, 1]$, with no graph text. Graph-reminder controls show the model its own earlier graph before the same prediction; replay controls reconstruct the earlier graph-report context before prediction.

Table 6. Representative coverage-faithfulness values. Coverage is a test-availability guardrail; cMPS and MaxShift are lower-better behavioral consistency errors; AUROC is an accuracy diagnostic. These are aggregate row summaries, so no standard deviation is reported.

Task	Model	T_g/T_p	Coverage	cMPS	MaxShift	AUROC
Admissions	Qwen 7B	0.2/0.0	0.714	0.024	0.354	0.731
Admissions	Qwen 7B	0.5/0.0	0.714	0.017	0.303	0.731
Loan default	Qwen2 7B	0.5/0.0	0.333	0.022	0.471	0.476
Loan default	Qwen2.5 14B	0.2/0.0	0.573	0.018	0.732	0.613
Loan default	Qwen2.5 14B	0.5/0.0	0.604	0.016	0.696	0.613
Loan default	Qwen 14B	0.5/0.0	0.722	0.002	0.062	0.496

Table 7. Exact role-conditioned estimates underlying Figure 2B. Excess shift is lower better. Means average graph-report-level role means so large raw test-count differences do not dominate; intervals are nonparametric 95% bootstrap confidence intervals over graph-report units.

Task	Graph role	Shift	95% CI	Cells	Graph units
Admissions	Ordinary irrelevant field	0.013	[0.012, 0.014]	46/96	344
Admissions	Demographic field marked irrelevant	0.015	[0.011, 0.019]	14/96	40
Admissions	Post-outcome leakage	0.031	[0.027, 0.034]	14/96	44
Loan default	Ordinary irrelevant field	0.012	[0.011, 0.012]	68/96	686
Loan default	Demographic field marked irrelevant	0.009	[0.008, 0.010]	60/96	474
Loan default	Post-outcome leakage	0.003	[0.000, 0.005]	5/96	5

Table 8. Exact aggregate graph-access controls behind Figure 3. ΔcMPS is relative to the fresh graph-absent baseline for the same task; positive values mean higher graph-contract violation. $|\Delta s|$ and Disagree measure score and decision movement relative to fresh on the same matched profiles. Each task/protocol row aggregates six model/task rows with licensed tests.

Task	Protocol	Cov.	cMPS	ΔcMPS	MaxShift	cFlip	$\lvert\Delta s\rvert$	Disagree	Pairs
Admissions	Fresh no graph	0.357	0.022	0.000	0.500	0.044	0.000	0.000	2000
Admissions	Graph reminder	0.357	0.035	+0.013	0.610	0.047	0.152	0.265	2000
Admissions	Replay	0.357	0.015	-0.007	0.300	0.023	0.091	0.152	2000
Admissions	Replay+graph	0.357	0.022	0.000	0.496	0.025	0.105	0.186	2000
Loan default	Fresh no graph	0.636	0.014	0.000	0.808	0.005	0.000	0.000	4400
Loan default	Graph reminder	0.636	0.034	+0.020	0.963	0.036	0.148	0.159	4400
Loan default	Replay	0.636	0.052	+0.038	0.974	0.071	0.269	0.497	4400
Loan default	Replay+graph	0.636	0.039	+0.025	0.899	0.060	0.286	0.416	4400