INTERVENTIONAL GROUNDING AUDITS: BLACK-BOX PREMISE-DEPENDENCY TESTS FOR LLM CHAIN-OF-THOUGHT VIA PREDICATE SUBSTITUTION

Hironao Nakamura
Independent Researcher

ABSTRACT

Large language models produce chain-of-thought (CoT) reasoning that appears logically sound yet may not genuinely depend on its stated premises. We introduce *interventional grounding audits*, a **black-box, step-level test of premise dependency**: we intervene on a single premise by **substituting its target predicate with a fresh symbol**, re-run the model, and check whether each reasoning step's **normalized conclusion** (canonical predicate form) changes. We evaluate on **ProntoQA**, a synthetic multi-hop deductive reasoning benchmark with gold proof trees, where step-level premise dependencies are known. Applied to 50 ProntoQA problems with GPT-4o, our method achieves F1 = 0.783 on detecting proof-tree dependencies (F1 = 0.835 on predicate-determining dependencies; Recall = 97.4%), significantly outperforming a self-consistency baseline (F1 = 0.346; 95% bootstrap CIs non-overlapping). We further identify that 28% of correctly-solved problems contain at least one step *insensitive* to proof-tree premises—a "right answer, wrong reasoning" phenomenon invisible to passive methods. All audit certificates, raw outputs, and reproduction scripts are included as supplementary material, and we discuss scope limits beyond formal, parsable benchmarks.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning can look logically sound while failing to genuinely depend on the premises it claims to use. CoT prompting enables large language models (LLMs) to solve multi-step reasoning problems by generating intermediate steps (Wei et al., 2022), but whether those steps are premise-dependent is often unclear. Consider an example from our experiments: GPT-4o solves a **ProntoQA** syllogistic problem correctly (a synthetic multi-hop deductive benchmark with gold proof trees, so each step's required premises are known), and self-consistency confirms this answer across five samples with 100% agreement. Yet our method reveals that four intermediate steps are *insensitive* to their proof-tree premises—the model reaches the right answer through reasoning that does not depend on its stated basis. This "right answer, wrong reasoning" (RAWR) phenomenon appears in **28% of correctly-solved problems**.

Passive methods—self-consistency (Wang et al., 2023), attention analysis—observe outputs but never intervene on inputs. They cannot distinguish genuine logical dependency from correlation. Self-consistency achieves only F1 = 0.346 on our dependency detection benchmark, primarily because it cannot identify *which* premises each step depends on (Precision = 0.230).

We propose *interventional grounding audits*: systematically substituting predicates in premises and observing whether each step's normalized conclusion changes. If replacing "tumpus" with an invented predicate "glumpus" in premise P_j changes step S_i's conclusion, S_i genuinely depends on P_j. This approach draws on a two-layer grounding framework whose operational implementation achieves state-of-the-art on agent safety benchmarks (Nakamura, 2026a;b), extended here from agent safety to reasoning integrity.

Our contributions: (1) An interventional protocol detecting premise-level causal dependencies with F1 = 0.783 (0.835 on predicate-determining dependencies), significantly outperforming self-

consistency (F1 = 0.346, non-overlapping CIs). (2) Two substitution strategies—consistent and local—with cascade filtering that distinguish direct from transitive dependencies (best F1 = 0.790). (3) A fully artifact-checkable evaluation: every certificate includes original and probed outputs with SHA256 checksums, verified by an automated validator.

2 METHOD

2.1 PROBLEM SETUP

Given premises $\{P_1, \ldots, P_k\}$ and a question, an LLM generates a CoT with steps $\{S_1, \ldots, S_n\}$. A proof tree specifies ground-truth dependencies: for each S_i, a set $\text{deps}(S_i) \subseteq \{P_1, \ldots, P_k, S_1, \ldots, S_{i-1}\}$ of **direct** (immediate-parent) dependencies; we do not take the transitive closure. Our task: for each pair (S_i, P_j), determine whether S_i genuinely depends on P_j. Section G discusses how 27% of false positives may be correct under a transitive definition. We model this through an *observation layer* (raw LLM text) and a *concept layer* (normalized propositions), intervening at the former and comparing at the latter.[1]

2.2 INTERVENTIONAL PROTOCOL

Predicate substitution (semantic probe). To test whether S_i depends on P_j, we replace the target predicate with an invented one (using a "zq" prefix). Two strategies: *consistent substitution* replaces the predicate in *all* premises, preserving chain coherence and detecting *predicate-determining* dependencies; *local substitution* replaces it *only* in P_j, breaking the chain and detecting *transitive* dependencies including structural premises. In ProntoQA, each premise introduces a unique predicate pair, so consistent predicate substitution is equivalent to premise-level intervention. In benchmarks with shared predicates across premises, this equivalence breaks down; local substitution addresses this case.

Surface rephrasing (control probe). We rephrase P_j without changing logical content ("All X are Y" $\rightarrow$ "Every X is a Y"). Output changes under surface rephrasing indicate surface sensitivity, not logical dependency.

Normalized proposition extraction. Each step's conclusion is parsed into a canonical form—$\text{is}(e, p)$ or $\text{subtype}(p_1, p_2)$—absorbing irrelevant variation in phrasing.

Five-value verdict. For each (S_i, P_j), comparing normalized conclusions under original (ϕ_{orig}), semantic probe (ϕ_{sem}), and surface probe (ϕ_{sur}): GROUNDED ($\phi_{\text{orig}} \neq \phi_{\text{sem}}$, $\phi_{\text{orig}} = \phi_{\text{sur}}$); INSENSITIVE (no change); INPUT-SENSITIVE (both change); UNSTABLE (only surface changes); UNPARSEABLE (parse failure). An INSENSITIVE step that explicitly cites P_j constitutes a **misrepresentation**.

2.3 CASCADE DETECTION AND FILTERING

Local substitution detects transitive dependencies but introduces cascade false positives: if local substitution on P_j changes S_i, all downstream steps also change via propagation. Our cascade filter reclassifies S_i as CASCADE if S_{i-1} is also GROUNDED w.r.t. the same P_j, recovering Precision ($0.604 \rightarrow 0.702$) while retaining Recall. This filter exploits the linear chain structure of ProntoQA; tree-structured proofs require generalization.

2.4 CONNECTION TO DEPLOYED GUARD ARCHITECTURE

Our pipeline mirrors a deployed safety guard achieving state-of-the-art on OS-level agent safety (Nakamura, 2026b): both use a two-stage *assess* $\rightarrow$ *decide* pipeline with named decision rules, structured verdict dataclasses, and automated evidence validators. Same architecture, different domain.

[1] This instantiates a formal grounding framework (Nakamura, 2025; 2026a); see Appendix E.

Table 1: Main results (GPT-4o, ProntoQA 50 problems). Bootstrap CIs confirm significance.

Method	P	R	F1	95% CI
Self-Consistency	0.230	0.698	0.346	[0.320, 0.371]
String-diff	0.738	0.803	0.769	—
A consistent (full)	0.735	0.838	0.783	[0.722, 0.842]
A+A′+cascade (full)	0.702	0.903	**0.790**	[0.716, 0.856]
A consistent (pred)	0.730	**0.974**	0.835	[0.760, 0.900]

Table 2: Ablation: contribution of each component.

Configuration	P	R	F1
A consistent only	0.735	0.838	0.783
A′ local only	0.644	0.899	0.751
A+A′ combined	0.604	0.908	0.726
A+A′ + cascade filter	0.702	0.903	**0.790**
A+A′ w/o surface ctrl	0.608	0.893	0.723
String-diff baseline	0.738	0.803	0.769

3 EXPERIMENTS

3.1 SETUP

We evaluate on ProntoQA (Saparov & He, 2023), 50 synthetic syllogistic problems (3–5 hops, 4–7 premises, True/False balanced). Target model: GPT-4o (temperature = 0). Dataset: 1,107 audit certificates. Parse rate: 90.7% (original), 86.9% (semantic probes); 213 UNPARSEABLE certificates (19.2%) are excluded from P/R/F1 computation. All counts are reported in the supplementary material. Metrics: P, R, F1 with 95% bootstrap CIs ($B = 10{,}000$, problem-level resampling). Two evaluation modes: F1_{full} (all proof-tree dependencies) and F1_{pred} (predicate-determining only).

3.2 MAIN RESULTS

Table 1 presents our main results with three key findings.

(1) Significant advantage over self-consistency. Our method (F1 = 0.783) significantly outperforms self-consistency (F1 = 0.346), with non-overlapping 95% CIs (gap = 0.351). The advantage is driven by Precision (0.735 vs. 0.230): self-consistency predicts *all* premises as dependencies for consistent steps, unable to identify which premises matter.

(2) Near-perfect recall on predicate-determining dependencies. Recall reaches 97.4% (4 misses out of 155) on predicate-determining dependencies, yielding $\text{F1}_{\text{pred}} = 0.835$. The 25 false negatives under F1_{full} (86%) are *structural premises*—entity-introduction premises (e.g., "Alex is a wumpus") whose predicate does not determine the step's conclusion. This is a granularity distinction, not a detection failure.

(3) Cascade filtering achieves highest F1. Combining consistent and local substitution with cascade filtering yields F1 = 0.790. Local substitution raises Recall (+0.065) by detecting transitive dependencies; cascade filtering recovers Precision lost to propagation FPs ($0.604 \rightarrow 0.702$).

3.3 ABLATION STUDY

Table 2 shows that consistent substitution alone achieves the best Precision (0.735) by preserving chain coherence. Local substitution adds Recall (+0.065) but reduces Precision (-0.131) from cascade FPs; the cascade filter recovers most of this loss. Surface control shows minimal effect on ProntoQA (ΔF1 = 0.003): our normalization to canonical is/subtype forms already absorbs the surface variation that rephrasing introduces, leaving little for the control probe to catch. We expect greater impact on natural language benchmarks where normalization is less effective.

3.4 ANALYSIS

Chain length. F1 decreases with chain length ($r = -0.62$), from 0.826 at 3 hops to 0.675 at 4 hops, driven by increased propagation FPs. Recall remains robust (>0.85) across all lengths (Appendix F).

False positives. Of 59 FPs (consistent sub.), 73% are stochastic output variation (addressable via majority voting), 27% are chain propagation effects arguably correct under broader dependency definitions (Appendix G).

RAWR (Right Answer, Wrong Reasoning). We define a problem as RAWR if: (i) the model's final answer is correct, and (ii) at least one step is rated INSENSITIVE to a direct proof-tree dependency. 14/50 problems (28%) satisfy both conditions. Self-consistency assigns perfect scores to all 14. Passive methods are blind to RAWR (Appendix C).

Misrepresentation. 12 cases where a step explicitly cites a premise yet is insensitive to it under substitution—CoT citation does not imply logical dependency (Appendix D).

4 RELATED WORK

Lanham et al. (2023) and Turpin et al. (2023) study whether CoT reflects true reasoning; we add a formal, per-step causal test. Wang et al. (2023) checks answer agreement; we show it cannot detect premise-level dependencies (F1 = 0.346). Vig et al. (2020) and Geiger et al. (2021) intervene on *internal* representations; we intervene on *inputs* (black-box, model-agnostic). Stolfo et al. (2023) apply causal interventions to mathematical reasoning; we extend to formal logical reasoning with a complete audit protocol. Saparov & He (2023) introduce ProntoQA; we add the interventional dimension, testing not just *whether* answers are correct but *whether each step depends on its stated basis*.

5 DISCUSSION AND LIMITATIONS

Limitations. ProntoQA is formal and synthetic; natural language benchmarks (FOLIO, ProofWriter) are needed to validate surface control and normalization components. Our results apply to *formal logical benchmarks with parsable steps*; generalization to free-form reasoning requires adapted normalization. Our dataset of 50 problems is modest, though bootstrap CIs confirm significance. Coverage matters: 77.7% of certificates receive definitive verdicts (22.3% UNPARSEABLE); treating all UNPARSEABLE as negatives yields a lower-bound F1 = 0.677 (vs. 0.783), still double the self-consistency baseline. Main results use GPT-4o only; Appendix B shows Claude Sonnet 4 exhibits qualitatively different probe responses (meta-reasoning about chain breaks), where coverage drops further—highlighting that parser robustness is a key deployment bottleneck.

Future work. Natural language benchmarks, majority voting ($k = 3$) for Precision improvement, and integration with the full formal framework (Nakamura, 2025; 2026a;c).

Reproducibility. All certificates, raw outputs, and scripts are available as supplementary material.[2] Every number is reproducible via the included scripts with no API key.

REFERENCES

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

[2] Supplementary evidence pack: https://github.com/hironao-nakamura/interventional-grounding-audits

Hironao Nakamura. Observation is a topos, not a type: A two-layer topos-HoTT framework for grounding and conceptual representation, 2025. URL https://zenodo.org/records/17894227. Preprint.

Hironao Nakamura. Grounded types as cross-layer invariants: Admissible updates and traceable witnesses in two-topos grounding, 2026a. URL https://zenodo.org/records/18253007. Preprint.

Hironao Nakamura. TTGOS Guard: Cross-benchmark safety SOTA. https://hironao-nakamura.github.io/ttgos-evidence/, 2026b. 100% on OS-Harm (NeurIPS 2025 Spotlight), CuP-SOTA on ST-WebAgentBench (ICLR 2026).

Hironao Nakamura. Interventional grounding: Identifiability and misrepresentation via counterfactual witnesses in two-topos grounding, 2026c. URL https://zenodo.org/records/18358924. Preprint.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.18653/v1/2023.acl-long.32. URL https://aclanthology.org/2023.acl-long.32/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

A Evidence Pack and Validation

Our full evidence pack is available at https://github.com/hironao-nakamura/interventional-grounding-audits and contains 4,486 files organized per-problem: 50 directories with original CoT outputs, all probe outputs (semantic, local, surface), and deterministic audit certificates for each (S_i, P_j) pair; self-consistency baseline data with five temperature-sampled runs per problem; and all source code with unit tests.

All file paths and source code are included; no external dependencies are required for Phase 2 (deterministic verdict computation) verification.

Verification: (1) SHA256 checksum file for integrity; (2) automated validation log for schema and referential integrity; (3) reproduction scripts to recompute all tables from raw certificates. The pipeline separates Phase 1 (LLM calls; non-deterministic) from Phase 2 (verdict computation; deterministic). Reviewers need only verify Phase 2.

B CROSS-MODEL ANALYSIS: CLAUDE SONNET 4

With the base normalizer, Claude Sonnet 4 achieves F1 = 0.161 (supplementary data), substantially below GPT-4o's 0.783. After adapting the normalizer for Claude-specific formatting (parenthetical asides, bracketed qualifiers), F1 improves to 0.320, still far below GPT-4o. When encountering inconsistent premises from substitution probes, Claude explicitly detects the inconsistency and produces *meta-reasoning* ("the chain breaks here because zqtumpuses are not mentioned…") rather than following the broken chain. These responses are functionally GROUNDED but unparseable into step-level propositions.

Example. Problem p038, Step 4 with P6 substituted: "*There is no premise that connects zqbrimpuses to any other category, so the chain of reasoning breaks here.*"

This reveals that different architectures exhibit qualitatively different strategies when encountering inconsistent premises—an observation uniquely enabled by our interventional approach. The supplementary evidence pack contains Claude certificates computed with the base normalizer; the 0.320 figure reflects a post-hoc normalizer adaptation not included in the pack. Adapting the normalizer to recognize meta-reasoning as grounding evidence is promising future work.

C RAWR CASE STUDY

Problem p026: GPT-4o correctly concludes "Knox is not a cralvus" with 100% self-consistency. Our audit identifies 4 of 5 steps as INSENSITIVE to at least one proof-tree dependency. Step 1 cites P4 but is insensitive to it; Steps 2–3 show similar insensitivity to structural premises. Self-consistency reports 100% confidence; our method reveals partially ungrounded reasoning.

D MISREPRESENTATION EXAMPLES

Problem p042, Step 1: "*Since Dove is a flompus, by premise 1, Dove is a flumpus.*" The step cites P6 ("Dove is a flompus"); under substitution (flompus $\rightarrow$ zqflompus), the conclusion remains $\mathrm{is}(\mathrm{dove}, \mathrm{flumpus})$. The step claims to depend on P6 but does not. We detect 12 such cases across 50 problems.

E FORMAL FRAMEWORK CONNECTION

Our two-layer model instantiates a formal framework (Nakamura, 2025; 2026a;c). The observation layer and concept layer are connected by $g^* : \mathrm{Obs} \rightarrow \mathrm{Concept}$ (normalize) and $g_* : \mathrm{Concept} \rightarrow \mathrm{Obs}$ (substitute). A GROUNDED certificate witnesses dependency via $g^* \dashv g_*$: intervening via g_* and observing via g^* reveals a change. An INSENSITIVE certificate witnesses absence. The complete certificate set provides a coverage witness analogous to the framework's completeness condition.

F CHAIN LENGTH ANALYSIS

Table 3: F1 by chain length (A consistent, full).

Hops	P	R	F1
3	0.735	0.943	0.826
4	0.559	0.851	0.675
5	0.602	0.933	0.732

Pearson $r = -0.62$ between chain length and F1. The dip at 4 hops is driven by increased propagation FPs, not Recall degradation ($>$0.85 at all lengths).

G FALSE POSITIVE BREAKDOWN

Of 59 false positives (A consistent only): 43 (73%) are *stochastic*—GPT-4o produces slightly different phrasing across runs even at temperature = 0. Majority voting ($k = 3$) should substantially reduce these. 16 (27%) are *chain propagation*—upstream substitution causes downstream changes. These are arguably correct under a transitive dependency definition.

H USE OF LARGE LANGUAGE MODELS

GPT-4o is used in this work solely as the experimental subject. For manuscript preparation, the author used Claude Opus 4.6 (Anthropic) and GPT-5.2 Thinking (OpenAI) as general-purpose assist tools for prose editing and document formatting. All research hypotheses, experimental design, method formulation, and scientific claims are the author's own.