Evaluating the Causal Effect of Chain-of-Thought on Groundedness in Tool-Use Agents via Counterfactual Mutations

Motivation. Reasoning-style language models improve tool-use agents, yet their visible chain-of-thought (CoT) may not always be faithful; steps can be decorative or even misleading while the final answer remains unchanged [1, 2]. We therefore target the causal effect of visible CoT on answer groundedness, i.e., whether articulating specific reasoning steps changes the probability that the final answer is supported by the fixed evidence, relative to answer-only or counterfactually edited CoT. This matters for systems that train from traces (distillation/SFT), deploy safety monitors that audit or shape rationales, and run tool agents that must follow an explicit plan: knowing whether the trace content *causally* improves groundedness tells us if visible traces are a reliable steering signal. [3].

Method. We propose a controlled evaluation that freezes the external context (retrieved passages and tool/API outputs) and compares four conditions for the same query: (A) *Baseline CoT* (model thinks then answers), (B) *No-External-Trace* ("answer-only" output schema; the model may reason internally but emits only the final answer + citations), (C) *Counterfactual CoT* where a pivotal reasoning step is automatically edited, and (D) *Counterfactual CoT + Grounding* which instructs the model to ignore any step that conflicts with evidence. We introduce a mutation library with labelable edits: *SalienceDrop*, *EntitySwap*, *Claim-AlignedDeletion*, *TopicDilution*, plus neutral controls (paraphrase/reorder). Judging is programmatic whenever possible (span/field entailment and contradiction checks); ambiguous cases are resolved with an LLM judge using majority vote. [4]

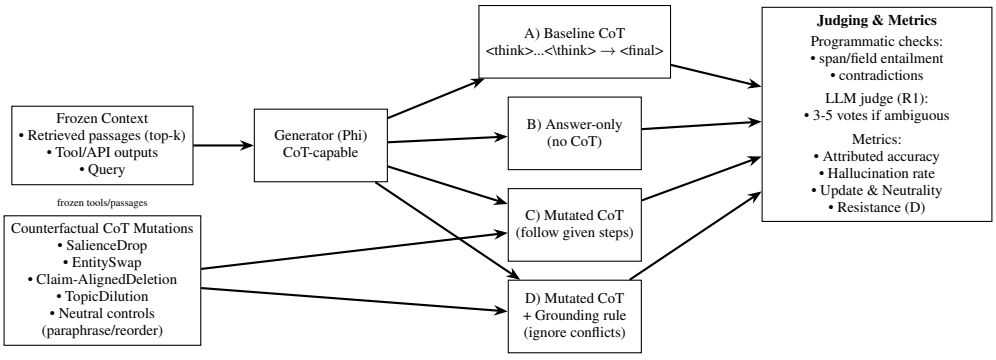
Metrics. Let $G \in \{0, 1\}$ denote whether the *final answer* is grounded (supported by the frozen evidence). We target the average causal effect (ACE) of visible CoT on groundedness:

$$\mathrm{ACE} = E[G \mid do(CoT = baseline)] - E[G \mid do(CoT = mutated)]\,.$$

With tools and retrieval frozen, we operationalize this via conditions A (Baseline CoT) and C (Mutated CoT), estimating $\widehat{\mathrm{CGE}} \approx \widehat{\mathrm{AA}}_A - \widehat{\mathrm{AA}}_C$, where $\widehat{\mathrm{AA}}_X$ is *Attributed Accuracy* in condition $X \in \{A, B, C, D\}$. We also report: $\Delta_{CoT \to AnsOnly} = \widehat{\mathrm{AA}}_A - \widehat{\mathrm{AA}}_B$, $\mathrm{Resistance} = \widehat{\mathrm{AA}}_D - \widehat{\mathrm{AA}}_C$, *Update Rate* (should-change edits), *Neutrality Rate* (control edits), *Hallucination Rate* (any unsupported claim), and token/latency budgets.

Expected Findings. A faithful CoT should (i) outperform answer-only generation on attributed accuracy, (ii) exhibit a high update rate for pivotal edits and high neutrality for controls, and (iii) recover under the grounding rule. Deviations expose decorative or brittle reasoning and help quantify the safety/efficiency trade-offs of CoT in tool-use agents; our framing connects to faithful-by-construction or plan-based approaches (e.g., SymbCoT, Faithful CoT) [6, 7].

Contributions. (1) A reproducible counterfactual-CoT benchmark for tool-use agents with frozen tools/retrieval; (2) an automatic mutation and judging pipeline; and (3) analysis guidelines for reporting faithfulness and groundedness under cost constraints. Code and scripts will be released to support double-blind, artifact-friendly reviewing.



References

[1] Turpin, M., Michael, J., Perez, E., Bowman, S. *Language Models Don't Always Say What They Think: Unfaithful Explanations in CoT*. NeurIPS 2023.

[2] Lanham, T. *et al. Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv:2307.13702, 2023.

[3] Korbak, T. *et al. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. arXiv:2507.11473, 2025.

[4] OpenAI. *PRM800K: A Process Supervision Dataset*. 2023.

[5] Manakul, A. *et al. Chain-of-Verification Reduces Hallucination in LLMs*. Findings of ACL 2024.

[6] Xu, J. *et al. Faithful Logical Reasoning via Symbolic Chain-of-Thought*. ACL 2024.

[7] Xie, V. *et al. Faithful Chain-of-Thought Reasoning*. IJCNLP-AACL 2023.