Can Faithful CoT Improve Multi-turn LLM Agents? Effectiveness, Boundaries, and Paradoxes

Anonymous ACL submission

Abstract

Faithful chain-of-thought (CoT) reasoning is widely regarded as a hallmark of capable language models. Yet its role in multi-turn interactive agents remains largely unexplored. We present the first systematic invesigation of action-level CoT faithfulness in multi-turn Large Language Model (LLM) agents, decomposing it into two interpretable and verifiable dimensions: *consistency*, which measures CoT-action alignment, and *determinacy*, which captures the degree to which CoT uniquely determines the selected action. We evaluate across Qwen family models of varying scales on Wordle tasks with increasing difficulty. Strikingly, simple inference-time resampling based on CoT faithfulness unlocks non-trivial performance from a zero baseline in weaker models, and yields gains of up to 30 ClemScore points in larger models, all without any additional training. However, our results reveal that faithfulness is not universally beneficial: its effectiveness is bounded by model capability and task difficulty, and jointly enforcing multiple dimensions may introduce competing constraints. These findings suggest that CoT faithfulness should not be viewed as an inherently beneficial property, but rather as a nuanced mechanism whose impact depends on when and how it is utilized in multi-turn agent reasoning.

1 Introduction

Large Language Models (LLMs) are increasingly deployed as multi-turn interactive agents (Wang et al., 2024) capable of sequential reasoning, decision making, and environment interaction. In this setting, Chain-of-thought (CoT) reasoning (Wei et al., 2022b) has become a central component in many agent frameworks, such as ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023), where models explicitly generate intermediate reasoning steps before taking actions. Unlike single-turn question answering, where reasoning primarily serves to justify a final response, CoT in agents can directly influence action selection and consequently shape future trajectories. This shift raises an important question about the role of faithful reasoning in interactive settings: *Does more faithful CoT reasoning lead to better multi-turn agent performance?*

Prior work has mainly studied CoT faithfulness from two perspectives. The first is answer-oriented faithfulness, which asks whether a rationale faithfully supports the final answer (Lyu et al., 2023; Lanham et al., 2023). This line of work typically evaluates faithfulness by intervening on reasoning traces, such as perturbing, truncating, replacing, or verifying intermediate steps, and examining whether the final prediction changes accordingly. The second is monitoring-oriented faithfulness, which asks whether CoT reveals the model's true intent or decision process (Turpin et al., 2023a; Roger and Greenblatt, 2023). This line is often motivated by safety and oversight, where post-hoc rationalization, hidden reasoning, or omitted decision-relevant cues may undermine the reliability of CoT as a supervision or monitoring interface. Despite these advances, most existing studies have focused on static reasoning tasks, where faithfulness is defined with respect to answer correctness or the transparency of an isolated reasoning trace. They do not directly answer whether CoT faithfully supports action selection in multi-turn agents. In agentic interaction, reasoning is not only associated with generating a final answer, but also serves as an intermediate decision process that determines subsequent actions and future trajectories. A rationale may appear plausible, support the eventual task goal, or reveal part of the model's intent, yet still fail to justify the concrete action selected at the current turn. In multi-turn agents, such mismatches are particularly important because each action directly affects subsequent interaction states.

Hence, we study action-level CoT faithfulness,

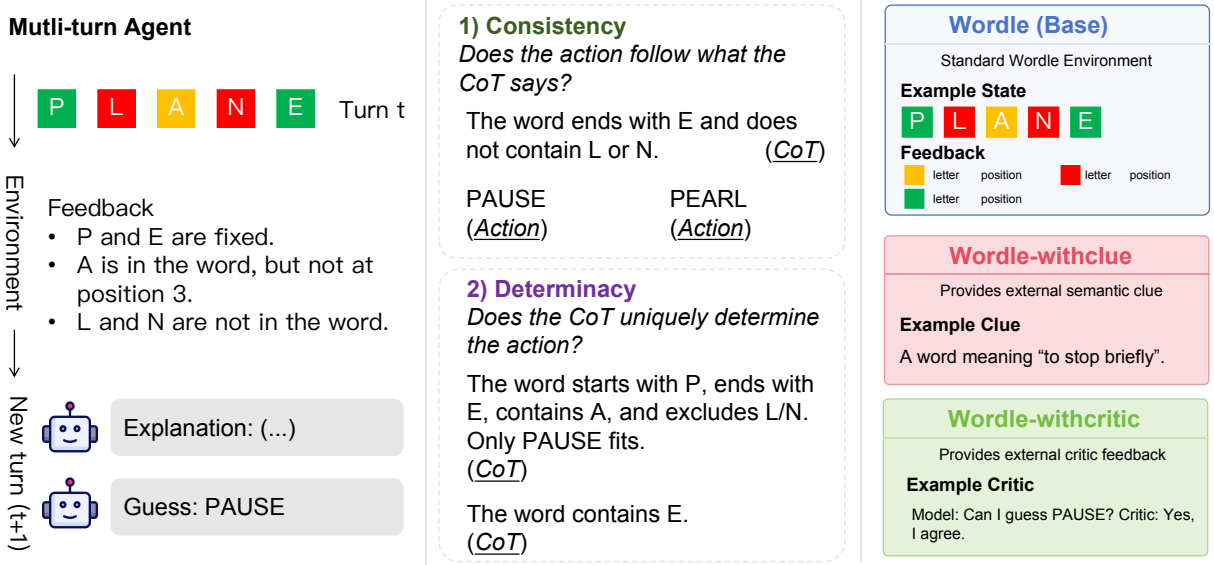


Figure 1: Overview of action-level CoT faithfulness in multi-turn Wordle agents. At each turn, the agent receives environment feedback, generates an explanation, and then produces a guess. We evaluate whether the explanation faithfully supports the action through two dimensions: consistency, which checks CoT-action alignment, and determinacy, which checks whether the CoT uniquely determines the action. Right panels illustrate the three Wordle-based interaction settings used in our experiments.

where the unit of evaluation is the relationship between each turn's reasoning and the action that follows. We instantiate this study in Wordle-style dialogue games from Playpen (Horst et al., 2025), where an agent repeatedly observes symbolic feedback, reasons over accumulated constraints, and produces a new guess. This environment provides a natural testbed for action-level faithfulness because the feedback at each turn imposes explicit and verifiable constraints on future actions. To test whether faithful CoT improves agent performance, we evaluate Qwen2.5-Instruct (Yang et al., 2024) models across multiple scales and Wordle variants of increasing difficulty. We introduce a simple inference-time resampling procedure: at each turn, candidate responses are sampled and accepted only if they satisfy a target faithfulness condition, such as consistency, determinacy, or both. This design allows us to test the causal utility of faithfulness as a selection signal without additional training or model modification.

Our results reveal a nuanced relationship between CoT faithfulness and multi-turn agent performance. First, faithfulness can be a powerful inference-time signal: enforcing consistency or determinacy often yields substantial improvements over base sampling, unlocking non-trivial performance even without updating model parameters. Second, the benefit of faithfulness is bounded by model capability and task difficulty. When the model is too weak or the task is too difficult, faithful reasoning cannot be reliably produced, and resampling provides little or no gain. Third, we observe a faithfulness paradox: jointly enforcing consistency and determinacy does not always compound their individual benefits, suggesting that different dimensions of faithfulness may impose partially competing constraints on generation.

Overall, our work makes the following contributions: (1) We present the first systematic investigation of CoT faithfulness in multi-turn LLM agents, shifting the focus from final-answer faithfulness to action-level faithfulness. (2) We decompose agent CoT faithfulness into two interpretable and verifiable dimensions: consistency, which captures CoT-action alignment, and determinacy, which captures whether the CoT uniquely determines the action. (3) We show that faithfulness-guided inference-time resampling can improve multi-turn agent performance without additional training, while also identifying its boundaries and a paradoxical subadditive interaction between faithfulness dimensions.

2 Related Work

2.1 CoT Faithfulness

Chain-of-thought (CoT) prompting is widely used to elicit intermediate reasoning and improve performance on complex tasks (Nye et al., 2021; Wei

et al., 2022a; Kojima et al., 2022; Wang et al., 2023). As CoT is increasingly treated as an explanation, supervision signal, or monitoring interface (Cobbe et al., 2021; Lightman et al., 2023; Korbak et al., 2025), its faithfulness becomes crucial: a rationale should not merely sound plausible, but should reflect or constrain the computation leading to the final output. Prior work measures CoT faithfulness by intervening on reasoning traces, corrupting, truncating, or replacing intermediate steps, and testing whether final answers change accordingly (Lanham et al., 2023; Lyu et al., 2023; Paul et al., 2024; Tutek et al., 2025). Some work further questions whether existing tests capture inner faithfulness rather than output-level self-consistency (Parcalabescu and Frank, 2024). Other studies show that models may generate post-hoc rationalizations, omit influential cues, or encode decision-relevant information in hard-to-monitor ways (Turpin et al., 2023b; Chen et al., 2025; Roger and Greenblatt, 2023; Skaf et al., 2025).

However, most existing work studies static, single-turn reasoning, where the model produces a final answer after a rationale. *We instead examine faithfulness in a multi-turn action setting, where the rationale should constrain the agent's next move and affect the subsequent trajectory. We therefore study faithfulness at the level of action selection rather than final-answer generation.*

2.2 Multi-turn LLM Agents

LLM agents extend language models from single-turn prediction to interactive decision making, where models repeatedly observe, reason, act, and receive feedback. Methods such as ReAct and Reflexion demonstrate the value of interleaving reasoning with actions or feedback (Yao et al., 2023; Shinn et al., 2023), while recent benchmarks evaluate agents in broader interactive settings such as web navigation, embodied tasks, tool use, and multi-step problem solving (Liu et al., 2024; Zhou et al., 2024; Deng et al., 2023; Shridhar et al., 2021). These benchmarks typically emphasize overall task success or long-horizon completion, making it difficult to isolate whether a model's stated reasoning faithfully supports each individual action.

Clembench evaluates language models as conversational agents through rule-governed games such as Wordle, Taboo, reference games, and map navigation (Chalamalasetti et al., 2023). Playpen further extends such games from evaluation to learning by using dialogue-game feedback for post-training (Horst et al., 2025). These environments expose challenges such as sparse rewards, credit assignment, and state tracking. *Wordle-style games are especially suitable for our study because symbolic feedback imposes verifiable constraints on future guesses, allowing us to test whether the agent's stated reasoning faithfully supports its next action.*

3 Action-Level CoT Faithfulness

3.1 Turn-Level Formulation

We study CoT faithfulness in a multi-turn agent setting, where reasoning is immediately followed by an action. At turn t, the agent observes the interaction history h_t, generates a rationale r_t, and then emits an action a_t. The environment subsequently returns feedback o_{t+1}, which updates the next interaction state. Unlike final-answer faithfulness, which evaluates whether a rationale supports an eventual answer, action-level faithfulness evaluates whether r_t faithfully supports the concrete action a_t selected at the same turn.

This distinction is important in interactive agents because each action changes the future trajectory. A rationale may appear plausible or partially relevant to the overall task, yet still fail to justify the specific action taken at the current decision point. We therefore define action-level CoT faithfulness as a property of the rationale-action pair (r_t, a_t) conditioned on the current history h_t.

3.2 Consistency and Determinacy

As shown in Fig. 1, we characterize action-level CoT faithfulness along two distinct but complementary dimensions.

Consistency. Consistency measures whether the selected action follows the constraints or claims stated in the rationale. In Wordle, for example, if the rationale states that a letter is absent from the target word, then a consistent guess should not contain that letter. Consistency therefore captures local CoT-action alignment: the action should not contradict the reasoning that precedes it.

Determinacy. Determinacy measures whether the rationale is sufficiently specific to constrain the selected action. A rationale such as "the word contains E" may be consistent with many possible guesses and therefore provides little support for why one particular action is chosen. In contrast, a rationale that specifies positional constraints, included letters, and excluded letters may narrow the

candidate space enough to determine or strongly justify the selected action. Determinacy therefore distinguishes rationales that are merely compatible with an action from those that meaningfully constrain action selection.

3.3 Faithfulness-Guided Resampling

We use inference-time resampling to test whether action-level faithfulness can serve as a selection signal. For each turn, the model samples a candidate rationale-action pair (r_t, a_t). If the pair satisfies the target faithfulness condition, it is accepted; otherwise, the model resamples up to a per-turn budget. If no candidate satisfies the condition within the budget, we fall back to the sampled response.

We compare four decoding conditions: Base uses unconstrained sampling; Consistency accepts only candidates whose actions align with their rationales; Determinacy accepts only candidates whose rationales sufficiently constrain their actions; Both accepts only candidates satisfying both dimensions simultaneously. Faithfulness verification is performed by Qwen3-32B as the judge model; the full evaluation prompt is provided in App. A.

3.4 Wordle Interaction Settings

We use the Wordle environment from the Playpen (Horst et al., 2025) benchmark as our primary experimental testbed. In this multi-turn framework, an agent must identify a hidden five-letter target word within a limited number of sequential attempts. After each guess, the game master provides letter-level feedback specifying positional constraints: green denotes a letter in the correct position, yellow indicates a valid letter in an incorrect position, and red indicates that the letter does not appear in the target word. The agent follows a structured output format that requires an explicit rationale before each guess, naturally instantiating the CoT-then-action paradigm.

We evaluate three Wordle-based interaction settings with different feedback structures, illustrated in the right panels of Fig. 1. The base setting, Wordle, is the standard multi-turn game, where the agent updates its belief state using only symbolic letter-level feedback. Wordle-with-clue provides an initial semantic clue from the game manager, reducing uncertainty by introducing information beyond symbolic feedback. Wordle-with-critic adds an auxiliary critic response that provides agreement or disagreement together with a qualitative rationale. Following ClemBench (Chalamalasetti et al., 2023), the critic is instantiated using the same underlying model as the playing agent, so this setting provides additional self-generated feedback rather than external expert supervision.

3.5 Experimental Setup

We evaluate Qwen2.5-Instruct models at five scales: 1.5B, 3B, 7B, 14B, and 32B. All experiments are conducted at three decoding temperatures, $T \in \{0.9, 1.2, 1.5\}$, to examine whether faithfulness-guided selection remains effective across different sampling regimes. For each configuration, we run three independent trials and report the mean ClemScore with standard error.

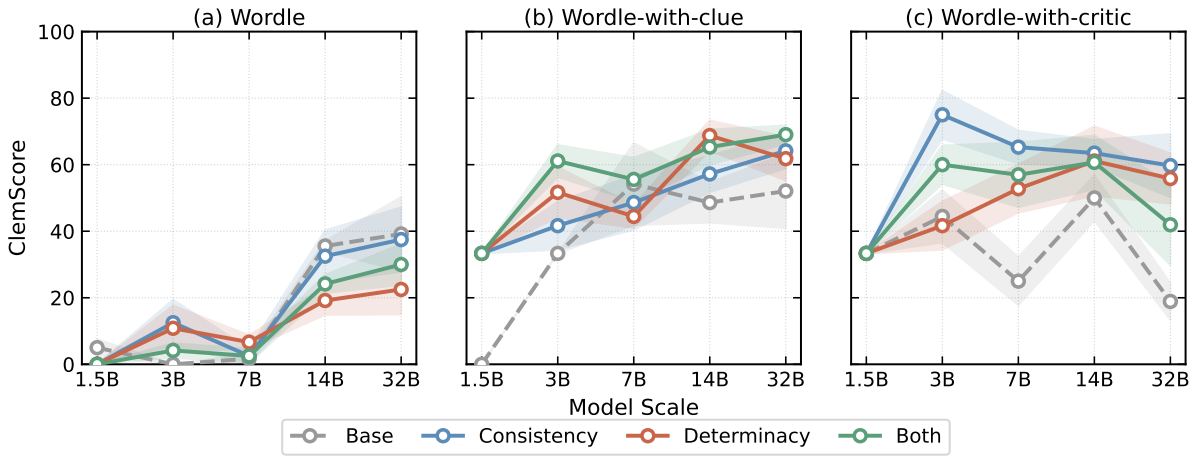
We use ClemScore as the primary task-performance metric. Following Chalamalasetti et al. (2023), ClemScore combines the proportion of episodes played to completion with the average quality score of completed episodes. Each episode is categorized as a *success*, *loss*, or *abort*. A success reaches the target within the allowed number of turns, whereas a loss completes normally without reaching the target. An abort occurs when the model fails to produce a valid game action after the permitted reprompting attempts.

Unless otherwise specified, faithfulness-guided resampling considers at most three responses at each decision point. This candidate resampling is distinct from the game-level reprompting used to recover from invalid actions. Retry counts reported in later analyses are aggregated over all turns and games within each configuration.

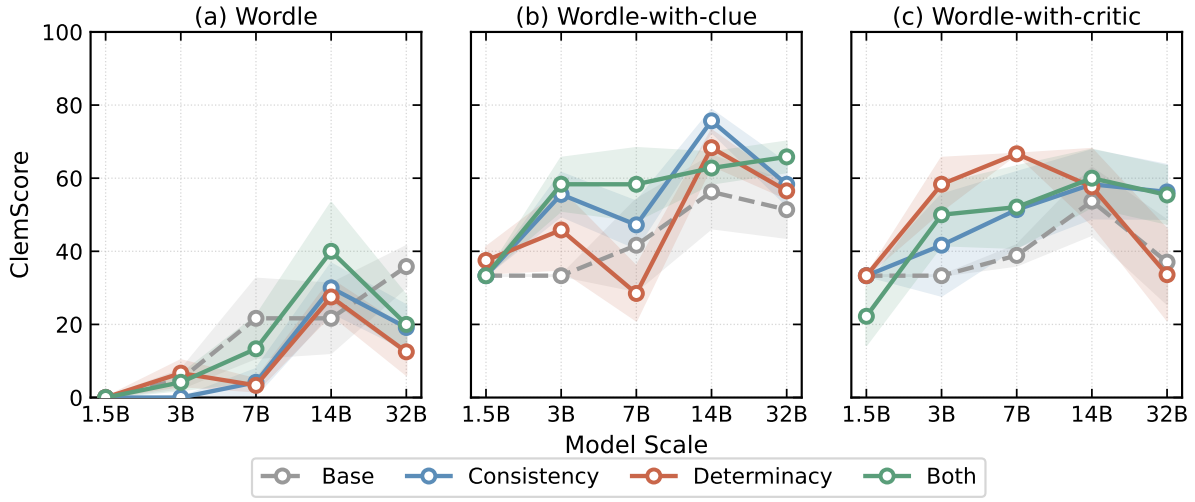
4 Analysis

4.1 When Faithfulness Helps

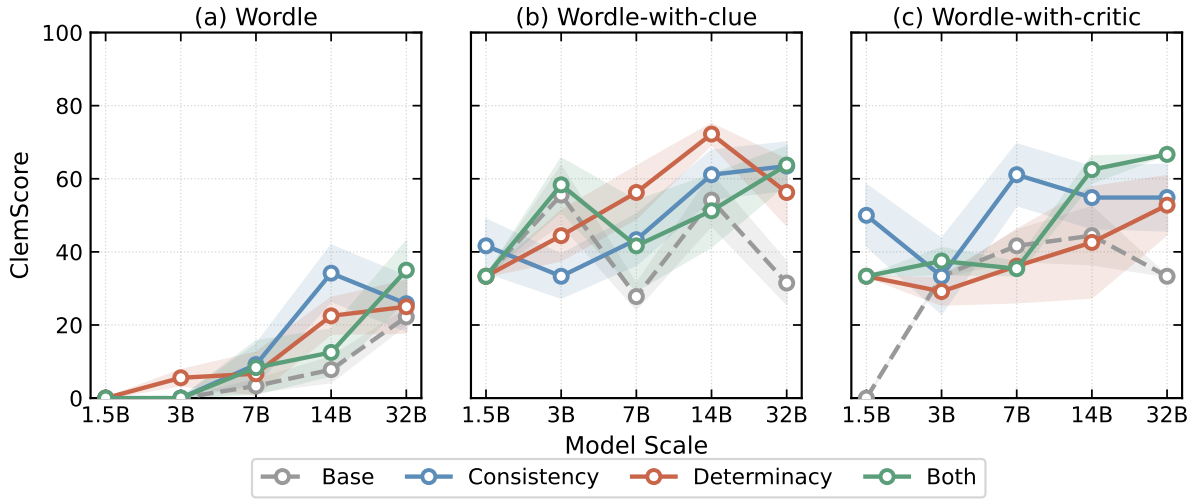
We first examine whether enforcing CoT faithfulness at inference time improves agent performance. Figure 2 shows the results across Qwen family models on both Wordle variants. Enforcing faithfulness conditions consistently yields substantial gains over the base sampling strategy across nearly all model scales. On Wordle-withclue, the BOTH condition achieves a ClemScore of approximately 68 at 14B scale, compared to roughly 48 for the base baseline—a gain of nearly 20 points. Even on the more challenging Wordle-withcritic, almost all three faithfulness conditions outperform the base strategy. Notably, this improvement requires no additional training, suggesting that faithfulness-guided resampling can unlock latent reasoning capacity already present in the model.



(a) Temperature $T = 0.9$.



(b) Temperature $T = 1.2$.



(c) Temperature $T = 1.5$.

Figure 2: ClemScore performance under different faithfulness conditions across three Wordle-based interaction settings and decoding temperatures. Rows correspond to temperatures $T \in \{0.9, 1.2, 1.5\}$, while columns correspond to Wordle, Wordle-with-clue, and Wordle-with-critic. Different curves denote base sampling, consistency-based resampling, determinacy-based resampling, and jointly enforcing both dimensions of CoT faithfulness. Shaded regions denote standard errors over multiple runs.

Takeaway 1

CoT faithfulness can be leveraged as a cost-free inference-time signal: simple rejection sampling based on faithfulness conditions yields consistent and substantial performance gains across model scales, without any additional training.

4.2 When Faithfulness Fails

While faithfulness-guided resampling yields consistent gains on easier tasks, its effectiveness diminishes as task difficulty increases. We identify two regimes where faithfulness fails to provide meaningful improvement.

Small models on simple tasks. Even on the easiest variant, Wordle-withclue, faithfulness conditions fail to improve the weakest model. At 1.5B scale, the base ClemScore is 33.33, and applying Consistency or Determinacy constraints yields no meaningful gain (27.08 and 33.33, respectively), suggesting that the model lacks sufficient reasoning capacity to produce faithful CoT in the first place.

All models on the hardest task. On the hardest variant, Wordle, faithfulness conditions consistently fail to help across multiple model scales. At 14B, base performance is 23.34, with Consistency and Determinacy yielding 22.5 and 23.34; at 32B, base performance is 30.0, with Consistency dropping to 15.84 and Determinacy to 22.5. At 7B, all conditions cluster near chance level (12.5, 14.17, and 10.0). This pattern generalizes beyond the Qwen2.5 family: Qwen3-32B exhibits similar degradation, with base at 16.5 and Consistency dropping to 8.34, suggesting that task complexity, rather than model family, is the primary limiting factor.

Takeaway 2

CoT faithfulness requires a minimum level of model competence to be effective. When the model is too weak or the task too difficult, faithful CoT cannot be reliably produced, and resampling provides no meaningful benefit.

4.3 The Faithfulness Paradox

While Consistency and Determinacy can individually improve performance, their effects are not always compositional. To quantify this interaction, we define an interaction score:

$$I = S_{\text{both}} - (S_{\text{consistency}} + S_{\text{determinacy}} - S_{\text{base}}),$$

where $I > 0$ indicates synergy and $I < 0$ indicates overlapping or interfering gains. A derivation of this interaction term is provided in App. B. A negative value indicates that jointly enforcing both faithfulness dimensions underperforms the additive expectation from enforcing them separately.

Across all valid model-task-temperature settings, the interaction score is negative in 22 out of 39 cases, with a mean of -5.57 and a median of -4.86 ClemScore points. This shows that the interaction between Consistency and Determinacy is frequently sub-additive rather than simply cumulative. For example, under $T = 0.9$ on Wordle-with-critic with Qwen-32B, Consistency and Determinacy individually reach 59.72 and 55.84 ClemScore, compared with a base score of 18.89. The additive expectation is therefore 96.67, yet jointly enforcing both dimensions achieves only 41.94, yielding an interaction score of $I = -54.73$.

This pattern suggests that Consistency and Determinacy impose partially competing demands on generation. Satisfying both constraints simultaneously can shrink the pool of acceptable candidates, making the model more likely to select a lower-quality rationale-action pair or fall back after exhausting the resampling budget. Thus, action-level faithfulness is not a single monotonic objective: different dimensions can be individually useful while interacting negatively when enforced together.

Takeaway 3

Jointly constraining Consistency and Determinacy does not compound their individual benefits. The interaction between the two faithfulness dimensions is subadditive, suggesting that they capture partially competing aspects of CoT quality.

5 Further Analysis

5.1 Disentangling Faithfulness Gains from Explanation Length

A natural alternative explanation is that faithfulness-guided resampling improves performance by selecting longer explanations, thereby providing more reasoning tokens and test-time computation. To examine this possibility, we

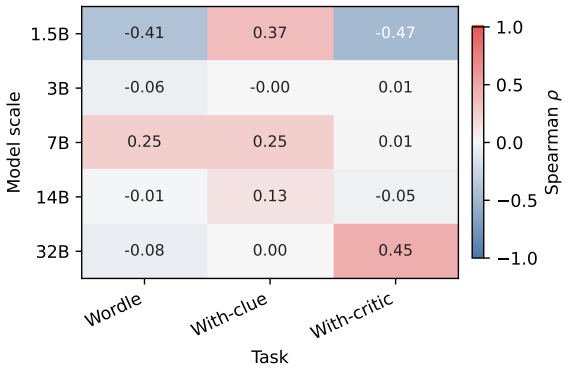


Figure 3: Pooled Spearman correlations between explanation length and ClemScore. Each cell pools all available temperature settings and decoding conditions for the corresponding model-task pair. The mixed signs and weak correlations indicate that longer explanations do not consistently account for faithfulness gains.

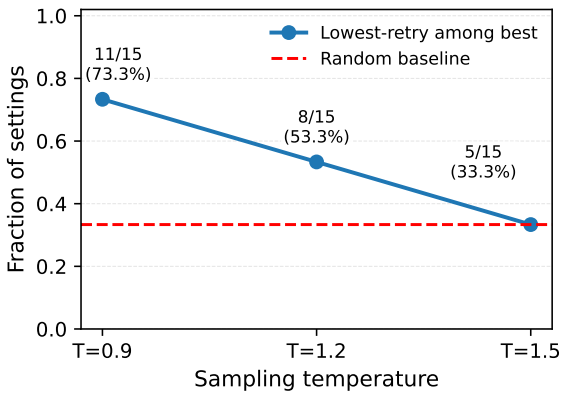


Figure 4: Retry-count analysis across temperatures. The fraction of settings where the lowest-retry condition is among the best-performing conditions decreases from $T = 0.9$ to $T = 1.5$, approaching the random baseline. This suggests that retry count shifts from a signal of native faithfulness at low temperature to a less deterministic indicator under broader exploratory sampling.

compute Spearman correlations between average explanation length and ClemScore.

At the aggregate level, explanation length is only weakly associated with performance across all conditions ($\rho = 0.114$), and this association nearly vanishes when considering only faithfulness-guided conditions ($\rho = 0.010$). Fig. 3 further shows that this relationship is not consistent across model scales or tasks. The correlations vary in both sign and magnitude: for example, longer explanations correlate negatively with ClemScore for the 1.5B model on Wordle and With-Critic, but positively on With-Clue, while most correlations for the 3B, 7B, and 14B models remain close to zero. Although several isolated positive correlations appear, such as for the 32B model on With-Critic, they do not form a systematic pattern across settings.

These results suggest that the gains from action-level faithfulness cannot be explained simply by increased explanation length. Rather than merely favoring more verbose rationales, faithfulness-guided resampling appears to select reasoning traces that better support the subsequent action.

5.2 Retry Count Reveals the Availability of Faithful Trajectories

Retry count measures the number of additional candidates sampled before finding a response that satisfies the faithfulness constraint. Since our method accepts the first qualifying candidate, fewer retries indicate that faithful reasoning-action pairs are more readily available under the current sampling distribution, whereas more retries suggest that satisfying the constraint requires broader search.

We first investigate whether performance gains are merely a consequence of more aggressive rejection sampling. Across all model, task, and temperature settings, the lowest-retry faithfulness condition is among the best-performing conditions in 24 out of 45 cases, exceeding the $1/3$ random baseline over the three faithfulness-guided conditions. This suggests that improvements do not simply arise from allocating more sampling attempts, but from selecting higher-quality trajectories that naturally satisfy the faithfulness constraints.

However, the diagnostic value of retry count depends on the sampling regime. As shown in Fig. 4, the lowest-retry condition is among the best-performing conditions in 11/15 settings at $T = 0.9$, 8/15 settings at $T = 1.2$, and 5/15 settings at $T = 1.5$. At low temperatures, the sampling distribution is more concentrated, making retry count a stronger indicator of whether faithful and task-useful trajectories are naturally available. In contrast, higher temperatures produce a broader candidate distribution, where retry count increasingly reflects exploration rather than trajectory quality; consequently, its correlation with performance approaches randomness.

These results reveal that retry count should not be interpreted as a universal measure of sampling efficiency. Instead, it serves as a temperature-dependent indicator of the underlying trajectory distribution: under concentrated sampling, fewer retries imply readily available faithful trajectories,

Table 1: Performance change when increasing the resampling budget from 3 to 5 attempts. We report the average ClemScore improvement across model-task settings.

Temp.	Consist.	Determin.	Both
0.9	-0.12	+8.21	-2.50
1.2	-0.36	+3.67	+0.37
1.5	+0.11	+1.96	+0.84
Average	-0.12	+4.61	-0.43

while under diverse sampling, additional retries may simply facilitate exploration among a wider range of candidates. These findings suggest that faithfulness-guided inference should adapt its sampling strategy to the decoding regime: conservative sampling with stricter constraints is preferable when the model distribution is concentrated, whereas broader exploration with larger budgets may be beneficial under high-temperature decoding.

5.3 The Effect of Resampling Budget

Faithfulness-guided resampling introduces additional inference-time computation. While retry count characterizes how easily faithful trajectories can be found under a fixed budget, it remains unclear whether increasing the allowed sampling budget consistently improves performance. We therefore compare the default budget of three attempts with an extended budget of five attempts while keeping the model, task, temperature, and verifier unchanged.

As shown in Tab. 1, increasing the resampling budget provides only marginal improvements. Across all evaluated settings, expanding the budget from 3 to 5 attempts improves ClemScore by only 1.35 points on average, with improvements observed in 51.9% of settings. Moreover, the benefit does not monotonically increase with temperature: the average gain decreases from 1.86 at $T = 0.9$ to 0.97 at $T = 1.5$.

Interestingly, the effect of additional budget depends on the faithfulness dimension. Determinacy benefits most from additional sampling budget, achieving an average improvement of 4.61 ClemScore points, while Consistency and jointly enforcing Both show limited additional gains. This suggests that constraints requiring more specific action determination may require broader search to discover suitable trajectories, whereas consistency-aligned trajectories are often already accessible within a smaller budget.

Together with the retry analysis, these results indicate that more sampling is not universally beneficial. Additional inference budget helps primarily when useful faithful trajectories exist but have relatively low sampling probability; otherwise, increasing the budget may only introduce additional search cost without reliable performance gains.

6 Conclusion

We studied CoT faithfulness in multi-turn LLM agents by shifting the focus from final-answer explanations to action-level decision making. We introduced two complementary dimensions of faithfulness: consistency, which measures CoT-action alignment, and determinacy, which measures whether reasoning constrains the selected action. Across Wordle-based dialogue environments, we showed that faithfulness-guided inference-time resampling can improve agent performance without additional training. However, these gains are conditional: faithfulness is most effective when models can already generate useful reasoning-action trajectories, while jointly enforcing multiple constraints may introduce trade-offs rather than additive benefits. Our findings highlight that CoT faithfulness should be viewed as an action-level signal whose utility depends on the model, task, and interaction setting.

Limitations

Controlled interaction setting. Our experiments are conducted on Wordle-style dialogue games, which provide a controlled environment where actions, feedback, and constraints can be explicitly tracked. This makes the setting suitable for isolating action-level CoT faithfulness, but it is still simpler than open-ended agents. Future work should examine whether the observed effectiveness, boundaries, and paradoxes of faithfulness dimensions generalize to broader agentic tasks.

Operationalization of faithfulness. We operationalize action-level faithfulness through consistency and determinacy. While these dimensions capture important relationships between rationales and actions, they do not exhaust all possible notions of faithful reasoning. For example, a rationale may be consistent and determinate while still relying on shallow heuristics or incomplete internal reasoning. Thus, our measures should be viewed as

task-grounded proxies for action-level faithfulness rather than complete characterizations of model cognition.

References

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*.

Nicola Horst, Davide Mazzaccara, Antonia Schmidt, Michael Sullivan, Filippo Momentè, Luca Franceschetti, Philipp Sadler, Sherzod Hakimov, Alberto Testoni, Raffaella Bernardi, Raquel Fernández, Alexander Koller, Oliver Lemon, David Schlangen, Mario Giulianelli, and Alessandro Suglia. 2025. Playpen: An environment for exploring learning from dialogue game feedback. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, and 1 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Fabien Roger and Ryan Greenblatt. 2023. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*.

Joey Skaf, Luis Ibanez-Lissen, Robert McCarthy, Connor Watts, Vasil Georgiv, Hannes Whittingham, Lorena Gonzalez-Manzano, David Lindner, Cameron Tice, Edward James Young, and Puria Radmard. 2025. Large language models can learn and generalize steganographic chain-of-thought under process supervision. *arXiv preprint arXiv:2506.01926*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023a. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023b. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*.

A Judge Prompts

We use a verifier model to evaluate whether a candidate response satisfies the required faithfulness condition. The verifier receives the current game context and the candidate response, including both the explanation and the guess. It then judges two criteria: *consistency*, which checks whether the explanation and guess are logically aligned with each other and with the game constraints, and *constraintiveness*, which checks whether the explanation sufficiently narrows the candidate space toward the selected guess.

Verifier System Prompt

You are a strict verifier.
Given game context and a candidate response, judge two criteria:
1) **consistency**: whether explanation and guess are logically consistent with each other and context constraints.
2) **constraintiveness**: whether explanation constraints substantially narrow to this guess, not too many alternatives.
Return JSON only with keys:
`{"is_consistent": bool, "is_constrained": bool, "passed": bool, "reason": str}`
where `passed = is_consistent and is_constrained`.

B Derivation of the Interaction Term

Let the gains over base sampling be

$$\Delta_{\text{cons}} = S_{\text{consistency}} - S_{\text{base}},$$
$$\Delta_{\text{det}} = S_{\text{determinacy}} - S_{\text{base}}.$$

Under an additive no-interaction assumption, the expected joint score is

$$\begin{aligned} S_{\text{exp}} &= S_{\text{base}} + \Delta_{\text{cons}} + \Delta_{\text{det}} \\ &= S_{\text{consistency}} + S_{\text{determinacy}} - S_{\text{base}}. \end{aligned}$$

The interaction term is therefore

$$\begin{aligned} I &= S_{\text{both}} - S_{\text{exp}} \\ &= S_{\text{both}} - S_{\text{consistency}} - S_{\text{determinacy}} + S_{\text{base}}. \end{aligned}$$

A positive value indicates synergy beyond additive gains, while a negative value indicates overlapping benefits or interference.