DRAFT DE-ANCHORING DECODING DOES NOT MITIGATE CONTEXTUAL DRAG IN LLM REASONING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Large language models in iterative reasoning workflows are susceptible to *contextual drag*—erroneous information in context biases subsequent generations even when the model is instructed to verify first. We propose Draft De-anchoring Decoding (D3), a training-free method that interpolates logits from draft-present and draft-absent key-value caches to attenuate harmful anchoring. Evaluating on Game of 24 with Qwen3-8B, we find that D3 fails both pre-registered success criteria: it does not improve wrong-draft accuracy (-0.65pp vs. required $+5$pp) and loses more than allowed on correct-draft accuracy (-2.21pp vs. allowed -1pp). Analysis reveals a fundamental mechanism flaw: draft-absent logits are too weak ($\sim$46% accuracy) to serve as a useful reference, and the method cannot distinguish beneficial from harmful anchoring. Our negative results suggest that decoding-time logit interpolation is insufficient for mitigating contextual drag.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in iterative reasoning workflows where they condition on previous outputs—prior attempts in refinement loops, tool outputs in agent systems, and retrieved passages in retrieval-augmented generation. In these settings, the model must use context selectively: benefiting from helpful information while avoiding being misled by erroneous content. Recent work has shown this is a major unsolved reliability problem. Cheng et al. (2026) introduced the phenomenon of *contextual drag*, demonstrating that conditioning on an incorrect draft solution can bias subsequent generations toward structurally similar mistakes, causing 10–20% accuracy drops even when the model is explicitly instructed to verify the draft first. Similarly, Lee et al. (2026) showed that reasoning models fail substantially when presented with contextual distractors.

Decoding-time logit manipulation has proven effective for related problems. Context-Aware Decoding (Xu, 2023) reduces hallucination by contrasting context-present and context-absent logits. Classifier-Free Guidance (Sanchez et al., 2023) improves controllability through conditional-unconditional interpolation. DoLa (Chuang et al., 2023) enhances factuality by contrasting logits across transformer layers. These successes suggest that a similar approach might mitigate contextual drag from erroneous drafts.

We propose Draft De-anchoring Decoding (D3), a training-free method that maintains dual key-value caches—one for the original draft-containing prompt and one for a draft-redacted prompt—and interpolates their logits at each decoding step. The hypothesis is that interpolating toward draft-absent logits will attenuate harmful anchoring when the draft is wrong while preserving utility when the draft is correct. We evaluate D3 on the Game of 24 task with Qwen3-8B using pre-registered success criteria.

Our contributions are as follows:

[1] https://gitlab.com/fars-a/draft-deanchoring-contextual-drag

- We confirm that contextual drag is a significant phenomenon: accuracy drops by 15.13 percentage points when conditioning on incorrect versus correct drafts, even with explicit verification instructions.
- We propose D3, a principled training-free approach based on logit interpolation, and rigorously evaluate it against pre-registered success criteria.
- We report that D3 fails both criteria: it does not improve wrong-draft accuracy (-0.65pp vs. required $+5$pp) and loses more than allowed on correct-draft accuracy (-2.21pp vs. allowed -1pp).
- We analyze why logit interpolation fails for this task: draft-absent logits are too weak ($\sim$46% accuracy), and the method cannot distinguish beneficial from harmful anchoring.

2 Related Work

Decoding-Time Logit Manipulation. Several methods modify output logits at inference time to improve generation quality without additional training. Context-Aware Decoding (Xu, 2023) reduces hallucination by contrasting logits from context-present and context-absent forward passes. Classifier-Free Guidance (Sanchez et al., 2023) interpolates between conditional and unconditional logits to improve controllability. DoLa (Chuang et al., 2023) contrasts logits from different transformer layers to improve factuality. DExperts (Liu et al., 2021) combines expert and anti-expert models for controlled generation, while Contrastive Search (Su & Collier, 2022) balances likelihood and diversity through token-level contrast. Our proposed D3 extends this paradigm to contextual drag mitigation by contrasting draft-present and draft-absent logits.

Contextual Influence in LLMs. Recent work has examined how context affects LLM behavior. Cheng et al. (2026) introduced the contextual drag phenomenon, showing that erroneous information in context biases subsequent reasoning. Liu et al. (2023) demonstrated that LLMs struggle to utilize information in the middle of long contexts. Zhou et al. (2024) studied robustness to noisy rationales in chain-of-thought prompting, while Douglas et al. (2024) proposed prior-aware decoding to mitigate distractor task influence. Lee et al. (2026) further showed that reasoning models fail with contextual distractors. Our work addresses contextual drag specifically in iterative refinement settings where models condition on potentially erroneous drafts.

Iterative Refinement. LLMs are increasingly used in iterative workflows where outputs are refined over multiple passes. Self-Refine (Madaan et al., 2023) enables models to iteratively improve their outputs through self-feedback. Reflexion (Shinn et al., 2023) uses verbal reinforcement learning for iterative improvement. Self-Consistency (Wang et al., 2022) samples multiple reasoning paths and selects the most consistent answer. Chain-of-Thought prompting (Wei et al., 2022) and its extensions such as Tree of Thoughts (Yao et al., 2023) and Plan-and-Solve (Wang et al., 2023) decompose complex reasoning into steps. These methods assume intermediate outputs are beneficial, but do not address the problem of erroneous intermediate outputs biasing subsequent generation.

3 Method

3.1 Problem Setup

We address the problem of *contextual drag* in iterative reasoning workflows (Cheng et al., 2026). Consider a setting where a language model is given a problem P along with a draft solution D in context, and is instructed to verify the draft before producing a final answer R. When D is incorrect, the model's output R is often biased toward the errors in D, even when explicitly instructed to verify first. This phenomenon, termed contextual drag, can cause substantial accuracy degradation compared to solving from scratch.

Formally, let $\mathcal{M}$ denote a language model and let $\text{Acc}(D_{\text{correct}})$ and $\text{Acc}(D_{\text{wrong}})$ denote the model's accuracy when conditioned on correct and incorrect drafts, respectively. The *contextual drag gap* is defined as:

$$\Delta_{\text{drag}} = \text{Acc}(D_{\text{correct}}) - \text{Acc}(D_{\text{wrong}})$$

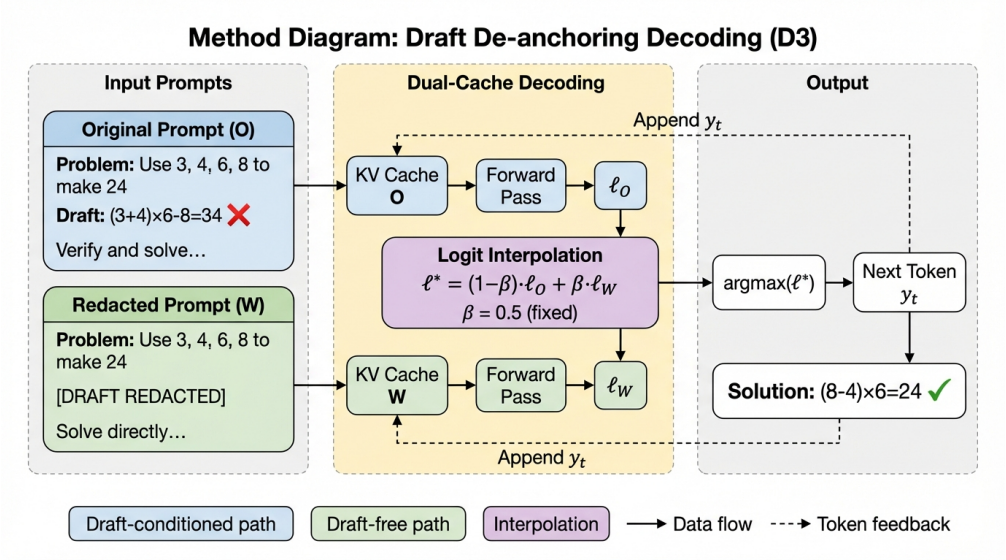

Figure 1: Overview of Draft De-anchoring Decoding (D3). The method maintains two parallel KV caches: one with the original draft-containing prompt (O) and one with a draft-redacted prompt (W). At each decoding step, logits from both caches are interpolated using an adaptive coefficient β scaled by Jensen-Shannon divergence, producing de-anchored logits for token selection.

Our goal is to reduce Δ_{drag} by improving $\text{Acc}(D_{\text{wrong}})$ while preserving $\text{Acc}(D_{\text{correct}})$—that is, to make the model more robust to erroneous context without sacrificing its ability to benefit from helpful context.

3.2 Draft De-anchoring Decoding (D3)

We propose Draft De-anchoring Decoding (D3), a training-free decoding method that interpolates logits from two parallel forward passes to reduce draft anchoring. As illustrated in Figure 1, the key idea is to maintain two key-value (KV) caches during generation: one for the original draft-containing prompt and one for a draft-redacted prompt. By interpolating toward the draft-absent logits, we aim to attenuate the influence of potentially erroneous draft content.

Prompt Definitions. Given an original prompt O containing the problem and draft solution, we construct a draft-redacted prompt W by replacing the draft span with a placeholder. Specifically, W instructs the model to solve the problem directly without reference to any draft, avoiding the contradictory instruction of verifying a nonexistent draft.

Decoding Procedure. At each decoding step t, we:

1. Compute logits $\ell_{O,t}$ from the original prompt O using its KV cache.
2. Compute logits $\ell_{W,t}$ from the redacted prompt W using its KV cache.
3. Compute the Jensen-Shannon divergence $\text{JSD}(\ell_{O,t}, \ell_{W,t})$ between the two distributions.
4. Calculate the effective interpolation coefficient: $\beta_{\text{eff}} = \beta_{\max} \cdot \min\left(1, \frac{\text{JSD}(\ell_{O,t}, \ell_{W,t})}{\tau}\right)$, where τ is a threshold parameter.
5. Interpolate: $\ell_t^* = (1 - \beta_{\text{eff}}) \cdot \ell_{O,t} + \beta_{\text{eff}} \cdot \ell_{W,t}$.
6. Select the next token: $y_t = \arg\max \ell_t^*$ (greedy decoding).
7. Append y_t to both sequences and update both KV caches.

The adaptive β scaling applies stronger de-anchoring only when the two distributions diverge substantially (indicating draft-induced bias), while preserving the original logits when they agree. We use $\beta_{\max} = 0.3$ and $\tau = 0.1$ based on hyperparameter optimization.

Greedy Decoding. We evaluate under greedy decoding (temperature = 0) to ensure that any performance changes result from genuine shifts in the argmax sequence, not from increased sampling diversity due to higher entropy.

3.3 HYPOTHESIS

The D3 method is motivated by the hypothesis that contextual drag operates through an anchoring mechanism in logit space. When the draft is incorrect, the draft-conditioned logits ℓ_O are biased toward draft-consistent tokens. By interpolating toward draft-absent logits ℓ_W, we hypothesize that D3 can reduce this bias and improve wrong-draft accuracy.

Crucially, we expect this intervention to be selective: when the draft is correct, ℓ_O and ℓ_W should produce similar distributions (since both lead to correct solutions), resulting in low JSD and minimal interpolation. When the draft is wrong, the distributions should diverge more, triggering stronger de-anchoring. This asymmetry would allow D3 to improve robustness to wrong drafts while preserving utility from correct drafts.

As we show in Section 4, this hypothesis does not hold in practice. The fundamental flaw is that the draft-absent logits ℓ_W are substantially weaker than the draft-present logits ℓ_O, and the method cannot distinguish between beneficial and harmful draft anchoring.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Task and Dataset. We evaluate on the Game of 24 task, where the goal is to construct an arithmetic expression using four given integers (each used exactly once) with basic operations ($+$, $-$, $\times$, $\div$) such that the result equals 24. We use 1,084 paired puzzles, where each puzzle has both a correct draft solution and an incorrect draft solution generated by the model. This paired design allows us to directly measure the contextual drag effect by comparing performance on the same puzzles with correct versus incorrect drafts.

Model and Decoding. We use Qwen3-8B (Yang et al., 2025) with greedy decoding (temperature = 0). For D3, we maintain dual KV caches and use adaptive β scaling with $\beta_{\max} = 0.3$ and JSD threshold $\tau = 0.1$. All experiments use the same prompt format from Cheng et al. (2026), which presents the problem and draft, then asks the model to verify the draft and produce a final answer.

Baselines. We compare against three baselines: (1) **1F Baseline**: Standard single-forward-pass decoding with the draft in context; (2) **Drop-Draft**: The draft is removed entirely from the prompt; (3) **Filler**: The draft is replaced with length-matched neutral filler text to control for context length effects.

Pre-registered Success Criteria. Following rigorous evaluation practices, we pre-registered two success criteria before running experiments:

- **Criterion 1 (Robustness)**: D3 should improve wrong-draft accuracy by at least 5 percentage points over the 1F baseline.
- **Criterion 2 (Utility Preservation)**: D3 should lose at most 1 percentage point on correct-draft accuracy compared to the 1F baseline.

4.2 MAIN RESULTS

Table 1 presents the main experimental results. We report four metrics: wrong-draft accuracy (accuracy when conditioned on incorrect drafts), correct-draft accuracy (accuracy when conditioned on correct drafts), mixed accuracy (50/50 weighted average), and the contextual drag gap (difference between correct-draft and wrong-draft accuracy).

Table 1: Main experimental results on Game of 24 with Qwen3-8B. D3 fails both pre-registered success criteria: it does not improve wrong-draft accuracy (Criterion 1: needed $\geq$+5pp, achieved -0.65pp) and loses more than allowed on correct-draft accuracy (Criterion 2: needed $\leq$1pp loss, achieved -2.21pp). Best results in **bold**. N/A indicates the metric is not applicable for that condition.

Method	Wrong-Draft Acc (%)	Correct-Draft Acc (%)	Mixed Acc (%)	Drag Gap (pp)
1F Baseline	**82.38**	**97.51**	**89.94**	15.13
Drop-Draft	46.03	N/A	N/A	N/A
Filler	49.08	48.34	48.71	N/A
D3 (original)	75.74	87.92	81.83	12.18
D3 (optimized)	81.73	95.30	88.51	**13.56**

Table 2: Effect of interpolation coefficient β on D3 performance. Higher β reduces the drag gap but also reduces overall accuracy. No β value achieves both improved wrong-draft accuracy and preserved correct-draft accuracy relative to the 1F baseline.

Configuration	Wrong-Draft Acc (%)	Correct-Draft Acc (%)	Mixed Acc (%)	Drag Gap (pp)
1F Baseline	**82.38**	**97.51**	**89.94**	15.13
$\beta = 0.1$ (fixed)	80.26	98.06	89.16	17.80
$\beta = 0.2$ (fixed)	81.18	96.22	88.70	15.04
$\beta = 0.3$ (adaptive)	81.73	95.30	88.51	**13.56**
$\beta = 0.5$ (adaptive)	76.94	91.88	84.41	14.94

Contextual Drag is Significant. The 1F baseline reveals a substantial contextual drag effect: accuracy drops by 15.13 percentage points when conditioning on incorrect drafts (82.38%) versus correct drafts (97.51%). This confirms that contextual drag is a real and significant phenomenon, even when the model is explicitly instructed to verify the draft before producing a final answer.

D3 Fails Both Success Criteria. The optimized D3 configuration fails to meet either pre-registered criterion. For Criterion 1 (robustness), D3 achieves 81.73% wrong-draft accuracy compared to the baseline's 82.38%, a decrease of 0.65 percentage points rather than the required 5-point improvement. For Criterion 2 (utility preservation), D3 achieves 95.30% correct-draft accuracy compared to the baseline's 97.51%, a loss of 2.21 percentage points that exceeds the allowed 1-point threshold.

Draft Content Drives the Effect. The control experiments reveal that contextual drag is driven by draft content, not context length or position. The Filler condition, which replaces the draft with length-matched neutral text, achieves only 48.71% mixed accuracy—similar to the Drop-Draft baseline (46.03%) and far below the 1F baseline with actual draft content (89.94%). This demonstrates that the model benefits substantially from draft content when it is correct, but this same mechanism causes harm when the draft is incorrect.

4.3 Beta Sensitivity Analysis

To understand whether the failure of D3 is due to suboptimal hyperparameter selection, we conduct a sensitivity analysis across different values of the interpolation coefficient β. Table 2 and Figure 2 present the results.

The results reveal a fundamental trade-off: increasing β reduces the contextual drag gap but also reduces overall accuracy. At $\beta = 0.1$, the method slightly improves correct-draft accuracy (98.06% vs 97.51%) but decreases wrong-draft accuracy (80.26% vs 82.38%), actually increasing the drag gap to 17.80pp. As β increases, both correct-draft and wrong-draft accuracy decrease, with correct-draft accuracy falling faster. The optimized $\beta = 0.3$ configuration achieves the smallest drag gap (13.56pp) but still fails both success criteria. Critically, no value of β achieves both improved wrong-draft accuracy and preserved correct-draft accuracy—the trade-off is monotonic with no sweet spot.

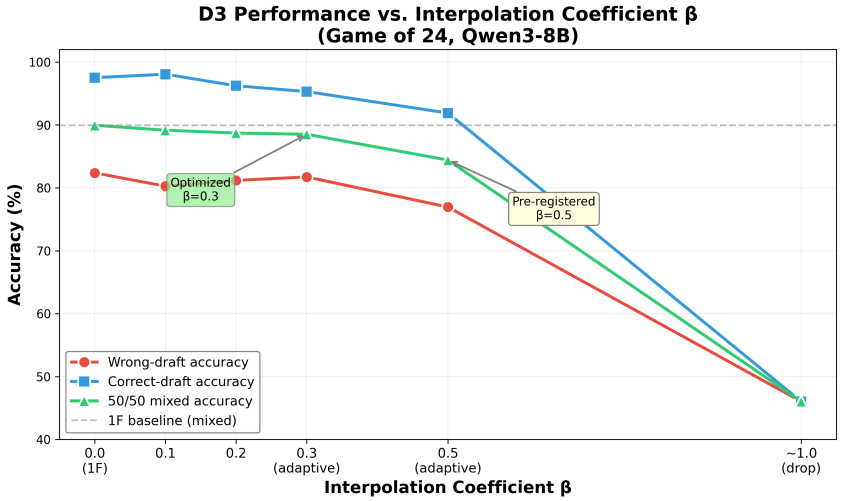

Figure 2: Effect of interpolation coefficient β on D3 performance. Higher β values reduce the contextual drag gap (difference between correct-draft and wrong-draft accuracy) but also reduce overall accuracy. The pre-registered $\beta = 0.5$ and optimized $\beta = 0.3$ configurations are annotated. No β value achieves both improved wrong-draft accuracy and preserved correct-draft accuracy relative to the 1F baseline.

4.4 Why D3 Fails

The failure of D3 can be attributed to a fundamental flaw in the logit interpolation mechanism. The method assumes that interpolating toward draft-absent logits ℓ_W will selectively reduce harmful anchoring while preserving beneficial anchoring. However, our analysis reveals three key issues.

Draft-Absent Logits Are Too Weak. The Drop-Draft baseline achieves only 46.03% accuracy, indicating that the draft-absent logits ℓ_W are substantially weaker than the draft-present logits ℓ_O (which achieve 89.94% mixed accuracy). Interpolating between strong ($\sim$90%) and weak ($\sim$46%) logits necessarily degrades overall performance. The method cannot "borrow" robustness from ℓ_W without also inheriting its lower accuracy.

Method Cannot Distinguish Draft Quality. The D3 hypothesis assumed that divergence between ℓ_O and ℓ_W would be higher for wrong drafts (indicating harmful anchoring) than for correct drafts (indicating beneficial anchoring). However, the average divergence rates are nearly identical: 5.57% for wrong-draft conditions and 6.10% for correct-draft conditions. This means the adaptive β scaling cannot distinguish between cases where de-anchoring would help versus hurt.

Interpolation Reduces Gap by Lowering Both. D3 does reduce the contextual drag gap (from 15.13pp to 13.56pp), but it achieves this by lowering both correct-draft and wrong-draft accuracy rather than selectively improving wrong-draft performance. The 10.4% relative reduction in the gap comes at the cost of 1.43 percentage points in mixed accuracy (89.94% $\rightarrow$ 88.51%). This is not a useful trade-off for practical applications.

5 Conclusion

We proposed Draft De-anchoring Decoding (D3), a training-free method to mitigate contextual drag by interpolating logits from draft-present and draft-absent forward passes. Despite its principled design, D3 fails both pre-registered success criteria: it does not improve wrong-draft accuracy and loses more than the allowed threshold on correct-draft accuracy. The fundamental flaw is that draft-absent logits are too weak ($\sim$46% accuracy) to serve as a useful reference, and the method cannot distinguish beneficial from harmful anchoring. Our negative results suggest that future work should

explore methods that can assess draft quality before deciding whether to de-anchor, or training-based approaches that learn when to rely on versus ignore contextual information.

REFERENCES

Yun Cheng, Xingyu Zhu, Haoyu Zhao, and Sanjeev Arora. Contextual drag: How errors in the context affect llm reasoning. 2026.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *ArXiv*, abs/2309.03883, 2023.

Raymond Douglas, Andis Draguns, and Tomáš Gavenčiak. Mitigating the influence of distractor tasks in lms with prior-aware decoding. 2024.

Seongyun Lee, Yongrae Jo, Minju Seo, Moontae Lee, and Minjoon Seo. Lost in the noise: How reasoning models fail with contextual distractors. 2026.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. pp. 6691–6706, 2021.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, F. Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, S. Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, A. Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651, 2023.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *ArXiv*, abs/2306.17806, 2023.

Noah Shinn, Federico Cassano, Beck Labash, A. Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. 2023.

Yixuan Su and Nigel Collier. Contrastive search is what you need for neural text generation. *ArXiv*, abs/2210.14140, 2022.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. pp. 2609–2634, 2023.

Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Zhichao Xu. Context-aware decoding reduces hallucination in query-focused summarization. *ArXiv*, abs/2312.14335, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, T. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601, 2023.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *ArXiv*, abs/2410.23856, 2024.