C^{3}D: Enhancing LLM Reasoning via Confidence-Guided Contrastive Decoding

Anonymous ACL submission

Abstract

Large language models (LLMs) are prone to distraction by contextual information during reasoning. Previous work primarily focuses on improving the generation of the next token while overlooking the potential bias introduced by existing premises. We propose a novel decoding method to mitigate such biases. Our framework uses predicted logits to estimate the model's confidence. By decomposing the full context into multiple premises, we gain a clearer understanding of the relevance of each premise to the question. During next-token prediction, we refine the output by contrasting the logits with the highest and lowest confidence. Our method effectively reveals how the model dynamically activates and adjusts its consideration of each premise as reasoning progresses.

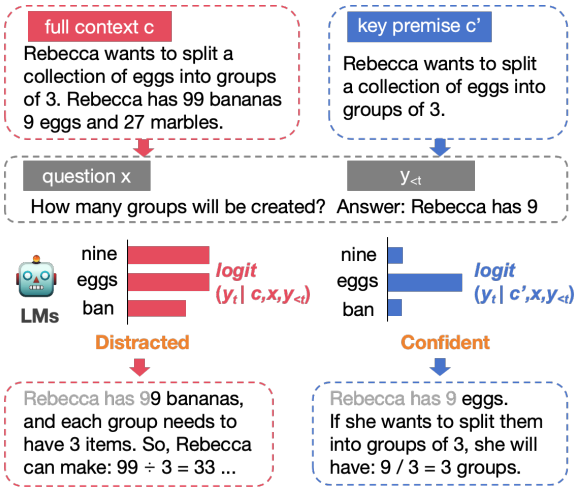


Figure 1: An illustration of a reasoning task. The language model becomes distracted by semantic coherence, thereby leading to error accumulation.

1 Introduction

Large language models (LLMs) have shown strong performance across a variety of reasoning tasks (Ahn et al., 2024; Zhang et al., 2024b; Plaat et al., 2024). As LLMs continue to advance, generating step-by-step rationale (Wei et al., 2022) helps guide them to more reliable answers. Many studies have focused on optimizing LLMs' reasoning process during inference. These efforts fall into two paradigms: *1) Fusion-based approaches*, which leverage information from the model itself or external sources to enhance reasoning robustness (Li et al., 2023; O'Brien and Lewis, 2023; Shi et al., 2024b). *2) Reasoning space search-based approaches*, which explore multiple reasoning paths to find the optimal solution (Wang and Zhou, 2024; Xie et al., 2023, 2024; Mo and Xin, 2024).

However, prior research has mainly focused on improving the generation of next tokens or rationales, while overlooking the influence of premises and context on subsequent outputs (Liu et al., 2024; Chen et al., 2024). Given that LLMs are autoregressive, **the preceding context profoundly shapes every subsequent token, including grammatical correctness, instruction adherence, and semantic coherence**. Yet, this dependency becomes a vulnerability in reasoning tasks. The intricate logical relationships within the context can cause the model to misinterpret cues or fixate on irrelevant details (Zhang et al., 2023a; Zhao et al., 2024; Chuang et al., 2023). This cognitive bias directly skews the token probability distribution, as illustrated in Figure 1. Such bias can lead to flawed reasoning sequences, with errors compounding over time.

A fundamental challenge in LLM reasoning stems from the opaque nature of its internal mechanisms (Malkin et al., 2022). The self-attention process, while powerful, makes it difficult to explicitly trace or control how the model conditions on specific premises. This lack of granular control often causes the model to default to prioritizing superficial signals, such as local syntactic coherence, over the logical integrity of the entire reasoning chain. **We argue that this sensitivity to contextual distractors is a primary driver of reasoning**

failures (Shi et al., 2023; Chen et al., 2024), an issue that proves highly resistant to correction via conventional, external interventions.

To address these challenges, we introduce Confidence-guided Cross-premise Contrastive Decoding (**C^{3}D**), a novel decoding strategy designed to mitigate distraction by actively guiding LLMs toward the most salient premises. Our approach is motivated by the key insight that LLMs reason more reliably with simple, explicit information (Zhou et al., 2022; Wang et al., 2023). Therefore, decomposing a complex context into a set of distinct premises can help the model better distinguish relevant from irrelevant information (Prystawski et al., 2023; Lightman et al., 2023). C^{3}D operationalizes this by first decomposing the problem's context into multiple premises. During generation, it then uses a confidence measure to dynamically identify the most and least supportive premise for the token being generated. Finally, it contrastively adjusts the next-token distribution by amplifying the influence of the highest-confidence premise while suppressing that of the lowest. This intervention effectively reduces reasoning biases arising from contextual ambiguity and fortifies the model against distraction.

We validate our method on multiple arithmetic and symbolic reasoning tasks. Experiments demonstrate that our method significantly improves performance without requiring training, external verifiers, or extensive search. Furthermore, our method enhances interpretability, offering deeper insights into LLMs' reasoning processes. In summary, our contributions are threefold:

- We propose a reasoning enhancement approach based on cross-premise awareness and contrastive decoding, introducing a token-level confidence evaluation to support the reliability of the model's reasoning chain.

- Our approach effectively reveals how language models dynamically activate consideration of different premises throughout the reasoning process. We further visualize each premise's influence on the generation of downstream tokens.

- The method achieves stable improvements in reasoning performance without requiring training, external verifiers, or path search. Extensive experiments validate its effectiveness.

2 Related Works

2.1 Large Language Models Reasoning

When confronting reasoning tasks, LLMs typically require CoT (Chain-of-Thought) (Wei et al., 2022) capabilities to perform step-by-step intermediate reasoning. Many studies focus on constructing more data to strengthen the underlying CoT abilities of LLMs, including methods based on Supervised Fine-Tuning (SFT) (Hao et al., 2024; Luo et al., 2023; Ranaldi and Freitas, 2024), Reinforcement Learning (RL) (Lightman et al., 2023; Zhang et al., 2024a), and Prompting techniques (Kojima et al., 2022; Zhang et al., 2022). These approaches alter the model's output logic and often demand high-quality data or evaluation models, as well as significant human effort and training costs.

2.2 Inference Time Scaling

In addition to training with more data, another technical approach explores improving LLMs during inference time. These methods aim to enhance the overall reasoning quality by designing effective supervision strategies for each step of the model's output, and it does not alter the model's inherent capabilities. Some studies employ internal or external auxiliary mechanisms to improve the robustness of LLMs (Li et al., 2023; Chang et al., 2023), while others opt for more direct approaches to search for optimal solutions within diverse reasoning spaces (Wang and Zhou, 2024; Xie et al., 2023, 2024; Mo and Xin, 2024). Our method falls into the category of internal model enhancement, which is low-dependency and low-overhead.

2.3 Contrastive Decoding

By contrasting a credible state with a non-credible state, contrastive decoding injects logits into the token generation process, thereby enhancing the faithfulness of the model's output from within (Shi et al., 2024a). For example, Contrastive Decoding (CD) (O'Brien and Lewis, 2023) uses an expert LM and an amateur LM to contrast and improve the professionalism of the generated tokens. Context-Aware Decoding (CAD) (Shi et al., 2024b), on the other hand, contrasts problems with and without context within a single LM to reduce the irrelevance of tokens to the context. Decoding by Contrasting Layers (DoLa) (Chuang et al., 2023) stimulates the intrinsic knowledge of LMs by contrasting different layers. COIECD (Yuan et al., 2024) utilizes information entropy to address the issue of

knowledge conflicts in models.

3 Method

We now introduce our proposed **C**onfidence-guided **C**ross-premise **C**ontrastive **D**ecoding (**C^{3}D**) method, which is a token-level, fine-grained premise-aware contrastive approach.

For a reasoning task, given an input question x and a context c that contains the necessary premises for reasoning, the generation process of a standard large language model $\mathcal{M}$ can be defined as:

$$\begin{aligned} y_t &\sim p_{\mathcal{M}}(y_t|c, x, y_{<t}) \\ &\propto \exp\big(\text{logit}_{\mathcal{M}}(y_t|c, x, y_{<t})\big) \end{aligned} \tag{1}$$

where y_t is the new token generated at time step t based on the context c, the question x, and the previously generated sequence $y_{<t}$. It is sampled proportionally to the logit scores processed by $\mathcal{M}$ (Shi et al., 2024b).

However, the default sampling method is influenced by various factors. For instance, when the context is complex or ambiguous, model predictions exhibit higher uncertainty (Zheng et al., 2023; Chen et al., 2024; Qiu and Miikkulainen, 2024), which manifests as a smoother logit distribution (Ulmer et al., 2023). This smooth distribution further leads to an averaging of sampling probabilities. Once the model selects an incorrect token, subsequent generations are affected as well. Even when the temperature is set to 0, it is difficult to guarantee that the top-ranked token is always correct. Moreover, to maintain linguistic coherence, the model tends to amplify these cumulative errors, ultimately compromising the correctness of the reasoning.

3.1 Confidence Estimation with Logits

To further explore the internal prediction mechanisms of the model, the logit lens (Belrose et al., 2023) has been employed for interpretability analysis. By inspecting the final-layer logits or probability distribution, we can observe how the model allocates probabilities across the vocabulary (Qiu et al., 2024; Yuan et al., 2024).

Generally, when a word is assigned a much higher probability than others, it indicates that the model has high confidence in the prediction and that it is more likely to be reliable (Zhang et al., 2023b; Duan et al., 2024). This often occurs with common collocations or when the intent is clear. Therefore, we employ the entropy of the predicted

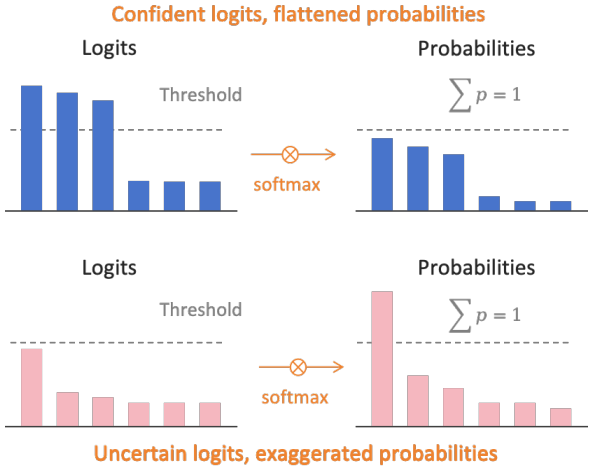


Figure 2: An example where entropy-based probability is insufficient to measure the model's confidence.

probabilities to measure the model's confidence α in the next token:

$$H = -\sum_{\tau \in \mathcal{V}} p_{\mathcal{M}}(\tau) \log\big(p_{\mathcal{M}}(\tau)\big) \tag{2}$$

$$\alpha(y_t) = \frac{1}{\exp(H_{y_t})} \tag{3}$$

where H is the entropy at the current position over the vocabulary $\mathcal{V}$. We further take the negative exponential of the entropy as an estimate of confidence. When the entropy is higher, the probability distribution over the vocabulary is more uniform, and the confidence is lower; when the entropy is lower, the distribution over the vocabulary becomes "sharper", and the confidence is higher (with a maximum value of 1).

However, entropy does not always faithfully represent the model's uncertainty, since some information is lost during the softmax process (Gupta et al., 2024; Ma et al., 2025). For example, the model may assign high logit values to multiple plausible words, but after softmax their probabilities are averaged out. Alternatively, when all logits are low, the model may still appear confident after softmax, which can artificially inflate the probability of a particular word, as illustrated in Figure 2. To address this limitation, we also take into account the extreme values of logits:

$$\mathcal{L}(y_t) = \frac{1}{K} \sum_{k=1}^{K} \text{topk}\big(\text{logit}_{\mathcal{M}}(y_t)\big) \tag{4}$$

where $\text{topk}(\cdot)$ extracts the largest k values from the logits. The idea behind this is that the magnitude of

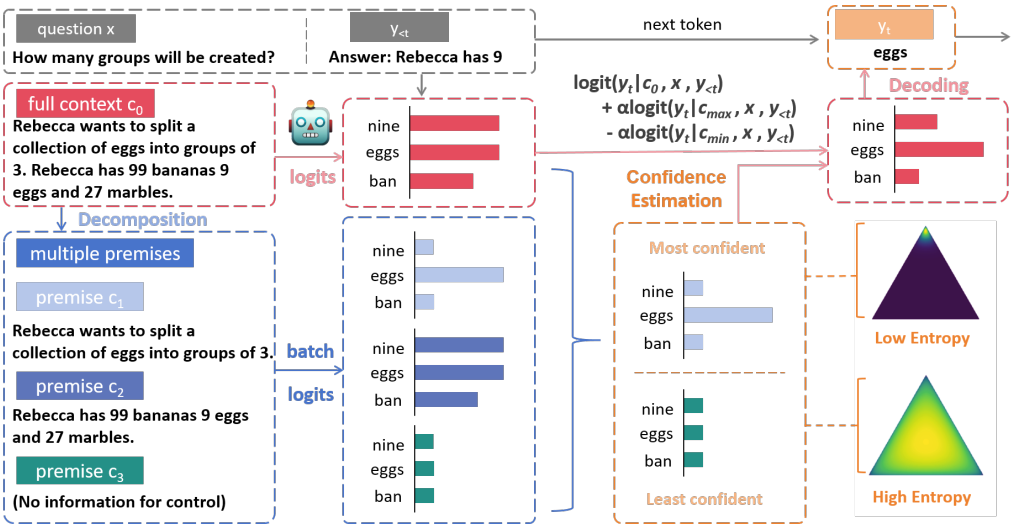


Figure 3: An illustration of our proposed C^{3}D method. The full context is decomposed into multiple premises, which then simultaneously obtain logits for the current position of the original question. By contrasting the most confident and least confident logits, the standard decoding process can be enhanced. This approach effectively mitigates the model's distraction issue. The illustration of entropy is copied from (Ulmer et al., 2023).

$\mathcal{L}$ can serve as an additional indicator of confidence (Ulmer et al., 2023). High logits indicate that the model's internal activation is strong, reflecting a high-energy state in which the model already exhibits clear preferences among candidate tokens. In such cases, applying the softmax transformation can unnecessarily flatten the distribution, obscuring the model's underlying certainty. Conversely, when logits are uniformly small, the model is in a low-energy state where differences among logits are subtle. In this regime, the softmax function plays a useful role by amplifying these small differences and providing a more informative probability distribution. By jointly considering the magnitude of logits and the resulting probability distribution, we can obtain a more faithful estimate of the model's confidence.

3.2 Multi-Premises Decomposition

Empirical observations suggest that models reason more effectively when tasks are expressed as simple, focused premises (Prystawski et al., 2023; Lightman et al., 2023). Such simplification reduces uncertainty and helps the model capture key information. Therefore, rather than treating the entire context c as a single unit, we decompose it into multiple premises $c_1, c_2, ..., c_n$, each representing a distinct and interpretable component of the reasoning problem. This decomposition makes the underlying logical structure explicit and provides a basis for contrastive decoding across premises.

Then, we can obtain the confidence level of each premise for the current position:

$$\alpha_n = \alpha(y_t|c_n, x, y_{<t}) \tag{5}$$

$$\mathcal{L}_n = \mathcal{L}(y_t|c_n, x, y_{<t}) \tag{6}$$

The hypothesis here is that when a premise is informative for the current decoding position, it will be assigned higher confidence. We aim to identify such premises and enable the model to distinguish the key information in the context from redundant details.

3.3 Dynamic Contrastive Decoding

To overcome reasoning errors caused by contextual distractions, we recompute the predicted logits during the decoding phase. Specifically, we select the premise logit with the highest confidence as the positive example and the premise logit with the lowest confidence as the negative example. We use their contrastive difference to adjust the original logits. Note that when the $\mathcal{L}$ values of all premises fall below a certain threshold, they are all considered untrustworthy, and in such cases, we rely solely on α as the confidence measure. Otherwise, we simply use $\mathcal{L}$ as our basis.

$$c_{max} = \begin{cases} \arg\max_{c_n}\{\mathcal{L}_0, \mathcal{L}_1, ..., \mathcal{L}_n\} \text{ if } \exists\mathcal{L} \geq T \\ \arg\max_{c_n}\{\alpha_0, \alpha_1, ..., \alpha_n\} \text{ if } \forall\mathcal{L} < T \end{cases} \tag{7}$$

$$\begin{aligned} \text{logit}'_{\mathcal{M}}(y_t|c, x, y_{<t}) &= \text{logit}_{\mathcal{M}}(y_t|c, x, y_{<t}) \\ &+ \alpha_{max}\text{logit}_{\mathcal{M}}(y_t|c_{max}, x, y_{<t}) \\ &- \alpha_{max}\text{logit}_{\mathcal{M}}(y_t|c_{min}, x, y_{<t}) \end{aligned} \tag{8}$$

where T is an empirically determined threshold, and $\mathcal{L}_0$ and α_0 denote the confidence of the full context. This decoding process is performed sequentially, and it dynamically selects a pair of contrastive examples for each generated token. Meanwhile, the confidence level α scales the magnitude of this adjustment. As a result, this method can mitigate the model's distraction by contextual information. Figure 3 presents the overall framework.

Algorithm 1 Confidence-guided Cross-premise Contrastive Decoding

Require: A reasoning task x with context c, and a language model $\mathcal{M}$
Ensure: Response sequence $y = \{y_1, y_2, ..., y_t\}$

```
Decompose c into premises {c_1, c_2, ..., c_n}
Add the full context and an empty set to the premise set C = {c, c_1, c_2, ..., c_n, ∅}
while t < max_length do
    Logit list ← ∅
    for c_i ∈ C do
        Add Logit_M(y_t | c_i, x, y_<t) to the Logit list
    end for
    if ∃L ≥ T for L in Logit list then
        Select c_max with the highest L and c_min with the lowest L
    else
        Select c_max with the highest α and c_min with the lowest α
    end if
    Contrast with c_max and c_min
    Sample y_t from the adjusted logits
    if y_t is eos_token then
        Break
    end if
end while
```

4 Experiments

We evaluate our method on multiple tasks that require models to reason based on context. We primarily focus on the following research questions:

- **RQ1**: (Performance) Can our method consistently improve reasoning performance?
- **RQ2**: (Case Study) How do multiple contextual premises influence the reasoning process?
- **RQ3**: (Mechanism) What is the relationship between the model's confidence and the downstream responses?

4.1 Experimental Setup

4.1.1 Language Models

To access internal logits, we conduct experiments on open-source large language models. Our base models are Llama-2-7B-chat and Llama-2-13B-chat (Touvron et al., 2023). Recently, specialized reasoning models, particularly those from the DeepSeek series (Guo et al., 2025), have shown exceptional performance. To test the generality of our approach, we additionally evaluate on these stronger reasoning models, specifically DeepSeek-R1-Distill-Llama-8B[1] and DeepSeek-R1-Distill-Qwen-7B[2].

4.1.2 Datasets

We validate our approach on commonly used benchmark datasets for reasoning, including three arithmetic reasoning tasks: GSM8K (Cobbe et al., 2021), AQuA (Ling et al., 2017), and SVAMP (Patel et al., 2021), as well as three symbolic reasoning tasks: Coin Flip (Wei et al., 2022), BIG-bench Date Understanding, and BIG-bench Object Tracking (Srivastava et al., 2023). Since the DeepSeek models already achieve strong performance on Coin Flip (exceeding 90%), we additionally evaluate on the AIME24 dataset (MAA, 2024) as a more challenging benchmark. These datasets encompass a wide range of reasoning tasks, from simple to complex, and require leveraging contextual information rather than relying on the model's memorized knowledge. Notably, the information provided in the questions is not always helpful, and some problems even contain completely irrelevant distractors. The model must carefully discern the given premises while avoiding reasoning pitfalls.

To validate the anti-distraction effect of our method, we also conduct tests on GSM-IC (Shi et al., 2023). This dataset is based on GSM8K but introduces irrelevant premises to the original questions, thereby distracting the language model. For experimental efficiency, we randomly sample 200 questions from GSM-IC as the test subset.

Since our primary focus is on how to make better use of the problem premises, we do not choose tasks like commonsense reasoning or mathematical computation. These tasks mainly rely on the model activating its stored knowledge for reasoning, where context information is usually minimal

[1] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

Models	Decoding	*Arithmetic*			*Symbolic*			Avg.	Throughput (tokens/s)
		GSM8K	**AQuA**	**SVAMP**	**Coin**	**Date**	**Object**		
Llama-2-7B-chat	Regular	21.68	24.01	41.90	47.00	39.29	30.80	34.11	48.41
	CAD	21.75	23.62	**49.90**	48.40	34.96	31.80	35.07	41.40
	DoLA	22.14	22.44	43.80	51.20	40.08	30.53	35.02	47.47
	Ours	**25.47**	**29.92**	47.59	**54.80**	**44.99**	**32.66**	**39.24**	45.70
Llama-2-13B-chat	Regular	34.49	15.74	49.40	47.40	46.07	27.33	36.84	30.26
	CAD	31.69	12.60	52.10	50.80	37.69	33.33	36.37	28.87
	DoLA	37.21	24.24	**55.70**	61.20	50.91	31.80	43.51	26.72
	Ours	**37.98**	**26.37**	55.10	**63.00**	**51.49**	**35.80**	**44.96**	27.54
		GSM8K	**AQuA**	**SVAMP**	**AIME24**	**Date**	**Object**	**Avg.**	
DeepSeek-R1-Distill -Llama-8B	Regular	76.47	75.59	87.20	30.00	76.15	81.33	71.12	33.96
	Beam	80.97	76.77	82.69	30.00	74.52	**88.40**	72.23	30.36
	CAD	81.80	79.13	**87.70**	33.33	77.23	81.46	73.44	29.89
	DoLA	82.25	77.55	87.10	30.00	74.79	81.86	72.26	25.79
	Ours	**83.85**	**81.10**	87.60	**36.67**	**85.90**	87.60	**77.12**	29.61
DeepSeek-R1-Distill -Qwen-7B	Regular	81.72	80.31	87.30	50.00	78.31	88.80	77.74	35.70
	Beam	80.67	83.25	88.80	46.67	**79.94**	88.40	77.95	34.07
	CAD	82.15	82.67	88.40	50.00	73.17	89.60	77.66	35.15
	DoLA	80.29	77.06	87.30	6.67	69.91	77.06	66.38	27.46
	Ours	**84.15**	**84.65**	**89.50**	**53.33**	79.40	**96.13**	**81.19**	34.03

Table 1: Performance (%) comparison across different decoding methods. Our proposed C^3D consistently improves performance across various arithmetic and symbolic reasoning tasks. Moreover, the enhancement effect of our method is more pronounced on stronger base models, such as DeepSeek-R1-Distill-Llama-8B.

or absent.

4.1.3 Baselines

We consider single-pass decoding methods as our baselines. Specifically, we compare with regular decoding, beam search decoding, context-aware decoding (CAD) (Shi et al., 2024b), and Decoding by Contrasting Layers (DoLA) (Chuang et al., 2023). Among these, CAD and DoLA are both contrastive decoding-based methods. The former primarily contrasts scenarios with and without context, while the latter focuses on contrasting different layers of the model.

4.1.4 Implementation Details

Our method introduces two hyperparameters: k, which controls the top k logit values for confidence $\mathcal{L}$, and threshold T, which serves as a threshold to balance the contributions of $\mathcal{L}$ and α. Specifically, k is searched over [1, 5, 10, 15, 20, 25] and T over [14, 16, 18, 20] (details can be found in Appendix A). Since our method requires simultaneous decoding across multiple premises, we employ a batched KV cache to improve efficiency. We perform all experiments on a single NVIDIA A800 GPU (80GB).

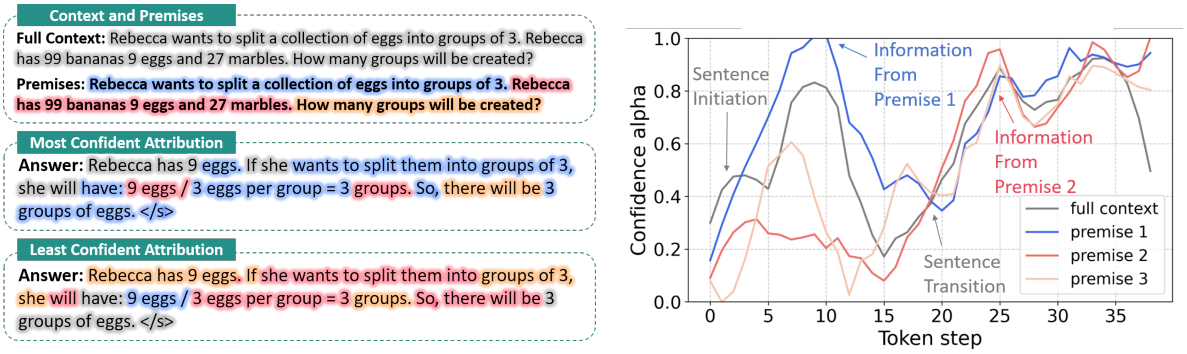
4.2 Overall Performance (RQ1)

Table 1 presents the performance of different models across various reasoning tasks. We further categorize the observations into Llama-2 Model Observations and DeepSeek-Distill Model Observations based on the reasoning capabilities of the models.

4.2.1 Llama-2 Model Observations

On the Llama-2 series, our method consistently improves regular decoding performance. Particularly on the AQuA and Coin Flip datasets, the 7B and 13B models show the most substantial improvements. AQuA contains non-intuitive and complex mathematical problems, while Coin Flip requires multi-step state tracking. Both tasks demand the model to thoroughly understand the problem. Given that the comprehensive understanding capability of the Llama-2 series is not particularly strong, the original decoding is easily influenced by the context. Our strategy, however, better assists the model in grasping finer-grained information.

For contrastive decoding methods, such as CAD and DoLA, their performance across different datasets is inconsistent. This suggests that relying solely on full-context contrast or layer-wise contrast is insufficient to obtain evidence for token generation.



(a) Visualization of predicted token attribution

(b) Confidence during reasoning flow

Figure 4: A visualized case study. Best viewed in color. The problem above in (a) is divided into three premises: we mark premise 1 in blue, premise 2 in pink, and premise 3 in orange, while the full context is marked in gray. The bottom boxes show which premise supports each generated token (most confident) and which distracts each generated token (least confident), respectively. The curve in (b) depicts how the confidence of each premise evolves as the generation progresses. The corresponding colors can help us better understand the reasoning process.

4.2.2 DeepSeek-Distill Model Observations

We further evaluate the performance of our method on stronger reasoning models. Although DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-7B already achieve strong results across multiple tasks, our approach yields additional gains in reasoning accuracy. Notably, the improvements are most pronounced on the Date Understanding and Object Tracking tasks, which likely benefit more from context-dependent information. At the same time, our method also provides stable improvements on standard arithmetic reasoning tasks. These findings suggest that stronger reasoning models are better able to leverage premise-based decoding, demonstrating both the scalability and robustness of our approach.

4.3 Performance with Irrelevant Context (RQ1)

Table 2 reports the performance comparison on the GSM-IC subset. GSM-IC introduces an irrelevant premise into each question, which often distracts the language model. Baseline models exhibit a significant performance drop compared to settings without irrelevant context, whereas our method largely preserves reasoning accuracy. These results demonstrate that our approach effectively reduces the adverse influence of irrelevant context during decoding.

4.4 Case Study (RQ2)

To gain a deeper understanding of how LLMs utilize known premises during the reasoning process, we further perform a case study for illustration.

Decoding	7B	13B	DS Llama	DS Qwen
Regular w/o IC	49.0	68.0	79.5	99.0
Regular w/ IC	35.0	55.0	76.0	95.0
CAD w/ IC	36.5	54.0	77.5	95.0
DoLA w/ IC	40.0	55.0	71.0	88.5
Ours w/ IC	41.5	62.5	83.0	97.0

Table 2: The performance (%) on the GSM-IC subset. With the insertion of irrelevant context into the questions, the baseline methods show significant performance degradation. Our method remains robust against such corruption.

Figure 4(a) shows the relationship between each premise in the problem and the downstream responses. We mark each premise with a distinct color and annotate the most confident and least confident premises for each generated token.

We observe that at the beginning of each response, the model relies more uniformly on the full context. It indicates that the global context still plays an important role in providing an overall foundation for reasoning. This ensures that potential dependencies spanning multiple premises are not lost after decomposition. As generation proceeds and more specific reasoning steps are required, the model shifts its attention toward the premises that provide the strongest support, such as premise 1. In contrast, premise 3, which contributes the least relevant information, initially shows the lowest confidence. Premise 2 also exerts some competing influence, leading to reduced confidence until later stages when its contribution becomes clearer.

Decoding	Llama2-7B		DS-Llama	
	GSM8K	AQuA	GSM8K	AQuA
C^3D	25.47	29.92	83.85	81.10
- w/o $\mathcal{L}$	19.11	28.35	70.25	74.80
- w/o α	23.09	28.74	82.97	78.74
Regular	21.68	24.04	76.47	75.59

Table 3: Ablation studies on $\mathcal{L}$ and α.

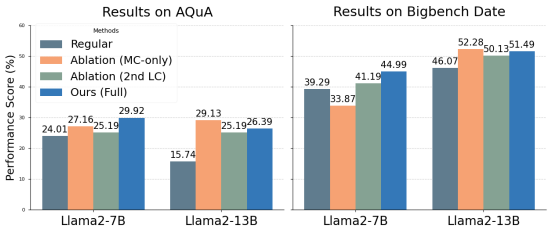


Figure 5: Ablation studies on contrastive components.

4.5 Understanding How Reasoning Flows (RQ2)

Figure 4(b) visualizes how the confidence values α of individual premises evolve during token generation. This analysis provides a clearer view of how the model directs the reasoning process. At the early stages of generation, premise 1 dominates, indicating that the model primarily relies on the most informative premise to initiate reasoning. As the generation progresses and intermediate information accumulates, initially less informative premises, such as premises 2 and 3, begin to contribute more actively, reflecting their growing relevance to the unfolding reasoning path. By the later stages, all premises converge to higher confidence levels, suggesting that each premise has been sufficiently integrated into the reasoning process. This dynamic progression highlights how our method reveals the shifting role of different premises during reasoning, and thereby provides a transparent account of how LLMs balance contextual evidence.

4.6 Impact of Confidence $\mathcal{L}$ and α (RQ3)

We validate the contributions of the two confidence measures, $\mathcal{L}$ and α, to reasoning performance. Table 3 presents the ablation studies on GSM8K and AQuA. The results show that both measures positively influence reasoning accuracy. However, removing $\mathcal{L}$ leads to a substantially larger performance drop compared to removing α. This highlights that relying solely on the entropy-based α is insufficient to capture the model's confidence. As discussed in Section 3.1, when the accumulated logits fall below a certain threshold, α becomes unreliable. The inclusion of $\mathcal{L}$ compensates for this limitation by directly reflecting the strength of the model's internal activations. It thereby plays a critical role in stabilizing the decoding process.

4.7 Analysis of Contrastive Mechanism (RQ3)

To understand the source of our method's performance, we conducted an ablation study dissecting its core mechanism (Figure 5). By testing a variant with only the most-confident distribution, we confirm that simply identifying a positive direction is insufficient. Furthermore, by comparing against a variant that subtracts a less extreme negative signal (the second-least confident distribution), we demonstrate that the precision of the negative guidance is critical. This confirms our central hypothesis: the most effective control is achieved by amplifying the most likely path while neutralizing the most significant distractor.

4.8 Analysis of Computational Overhead

To quantify our method's computational overhead, we evaluated its decoding throughput (tokens per second) against baselines. The results in Table 1 indicate that our method's performance is on par with widely-used strategies like CAD and DoLA. Our implementation utilizes a straightforward approach where each of the n paths maintains a separate KV cache. The observed efficiency stems from the fact that the dominant computational cost in generation is the model's forward pass, which is composed of highly parallelizable matrix operations. Consequently, our method achieves its significant quality gains while maintaining a computational footprint comparable to existing advanced decoding algorithms, making it a practical solution.

5 Conclusion

We propose a confidence-guided cross-premise contrastive decoding method that effectively mitigates reasoning biases in LLMs arising from contextual distractions. We validate the method on both weaker reasoning models and recent state-of-the-art reasoning models. Experiments show that our approach yields consistent gains. Furthermore, by analyzing and visualizing the dynamic flow of reasoning across premises, our method offers new insights into how LLMs process context, which can inform the design of future reasoning-oriented architectures and strategies.

Limitations

Our approach, like all methods that rely on an LLM's output probabilities, is subject to the calibration of the base model. Consequently, in scenarios where the model is "confidently wrong," the effectiveness of our confidence-based guidance mechanism can be constrained. It is important to note that this is not a limitation unique to our method, but an inherent challenge for any technique building upon current-generation LLMs (Wan et al., 2024; Taubenfeld et al., 2025; Agarwal et al., 2025). Our experiments demonstrate a consistent improvement over baselines that operate under the same constraint. Therefore, a promising direction is to couple our decoding strategy with future advancements in model calibration and uncertainty quantification.

Furthermore, our approach presupposes that the provided context can be effectively segmented into distinct, sentence-level semantic units. This assumption holds true for a wide range of tasks involving well-structured prose, such as the reasoning tasks we evaluated. However, its performance may be different on texts that are not easily divisible or where the semantic unit is not a sentence, such as in poetry, legal documents with long convoluted clauses, or source code. Adapting our method to operate on different granularities, such as automatically identified "conceptual chunks" rather than sentences, is a non-trivial but important direction for future research to broaden its applicability.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. 2023. Kl-divergence guided temperature sampling. *arXiv preprint arXiv:2306.01286*.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

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

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. *Preprint*, arXiv:2110.14168.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*.

Shibo Hao, Sainbayar Sukhbaatar, D Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space, 2024. *URL https://arxiv.org/abs/2412.06769*.

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*, volume 35, pages 22199–22213. Curran Associates, Inc.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

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*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(ACL), pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. 2025. Estimating llm uncertainty with logits. *arXiv preprint arXiv:2502.00290*.

MAA. 2024. Maa invitational competitions.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

Shentong Mo and Miao Xin. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12742–12746. IEEE.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

Ben Prystawski, Michael Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Advances in Neural Information Processing Systems*, volume 36, pages 70926–70947. Curran Associates, Inc.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024a. A thorough examination of decoding methods in the era of llms. *arXiv preprint arXiv:2402.06925*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20090–20111, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2023. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2024. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. *arXiv preprint arXiv:2408.17017*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*, volume 37, pages 66383–66409. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. In *Advances in Neural Information Processing Systems*, volume 36, pages 41618–41650. Curran Associates, Inc.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Rest-mcts : Llm self-training via process reward guided tree search. In *Advances in Neural Information Processing Systems*, volume 37, pages 64735–64772. Curran Associates, Inc.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024b. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

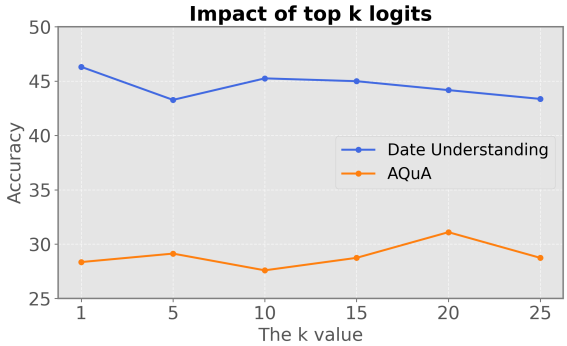


Figure 6: The trend of accuracy impact under different top-k values.

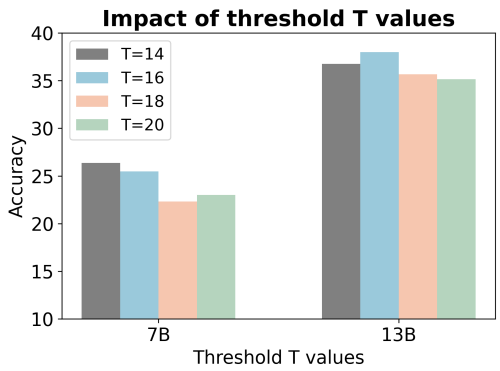


Figure 7: The performance of different T values on GSM8K across various models.

A Hyperparameter Analysis

A.1 Impact of Hyperparameter k

Figure 6 illustrates the performance trends of AQuA and Date Understanding under different top-k logit settings. As k increases from 1 to 25, the curves exhibit different trends. This divergence reflects the intrinsic differences between tasks: Date Understanding relies on a narrow set of date-related tokens, while AQuA involves a broader lexical space. Thus, selecting a moderate k offers a balanced compromise, ensuring sufficient coverage for tasks with diverse vocabularies while avoiding performance degradation in tasks with narrower token distributions.

A.2 Impact of Hyperparameter T

The role of logit extremal values in shaping responses has received limited attention, making it challenging to define a reasonable threshold. To investigate their effect, we empirically explore thresholds in the set [14,16,18,20]. Figure 7 illustrates the impact of different thresholds T on GSM8K across two models. We observe that, despite variations in model size, the range of logits remains consistent and their influence is largely similar. This finding suggests that selecting a moderate threshold provides a stable reference across datasets and model scales.

B Case Study on GSM-IC

To further analyze how distraction affects the model's reasoning process, we compare the performance of different decoding methods on the GSM-IC dataset, as shown in Table 4. We observe that regular decoding is more susceptible to irrelevant information in the questions, leading to incorrect responses. In contrast, our method empowered by cross-premise contrastive decoding can selectively ignore such distracting information.

Regular	**C^{3}D**
Question: Gretchen draws caricatures in the park on the weekends. She charges $20.00 per drawing. Ideally, Emma would love to sell 10000 drawings per weekend. If Gretchen sold 24 on Saturday and 16 on Sunday, how much money did she make?	**Question:** Gretchen draws caricatures in the park on the weekends. She charges $20.00 per drawing. Ideally, Emma would love to sell 10000 drawings per weekend. If Gretchen sold 24 on Saturday and 16 on Sunday, how much money did she make?
Answer: Gretchen sells 24 drawings on Saturday and 16 drawings on Sunday. So, in total, she sells 24 + 16 = 40 drawings. Since each drawing costs $20.00, Gretchen makes $20.00 x 40 = $800.00 per weekend. To find out how much money Emma would love to sell, we need to multiply the number of drawings Gretchen sold by the price per drawing: $800.00 x 10000 = $8,00,000.00. (Wrong)	**Answer:** Gretchen charges $20.00 per drawing, so if she sells 24 drawings on Saturday, she will make $20 x 24 = $480.00. If she sells 16 drawings on Sunday, she will make $20 x 16 = $320.00. So, in total, Gretchen made $480.00 + $320.00 = $800.00. (Correct)

Table 4: A comparative case study of different decoding methods on the GSM-IC dataset. Red text highlights distracting irrelevant information that negatively influences the model. The left panel demonstrates how standard decoding fails when exposed to such distractions, while the right panel shows our method's robust performance under the same conditions.