Ordering Matters: Enhancing LLM Logical Reasoning via CoT-Guided Context Reordering

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) excel at diverse NLP tasks, they struggle with complex logical reasoning due to an unaddressed vulnerability to disordered input contexts. We argue that this stems from a fundamental structural misalignment: logical deduction inherently forms a non-linear graph, whereas decoder-only LLMs rely on causal attention, enforcing a strict left-to-right sequence, which structurally blinds them to trailing prerequisites. To bridge this gap, we introduce CoT-Guided Context Reordering (**CGCR**), a training-free, inference-time framework that physically aligns logical dependencies with the model's linear attention flow. CGCR leverages the LLM's internal reasoning to generate a structured Chain-of-Thought (CoT), extracts premise-level causal dependencies to map a Directed Acyclic Graph (DAG), and applies topological sorting to prioritize core proof chains while relegating irrelevant noise to the end. Extensive experiments across three logical benchmarks demonstrate that CGCR consistently and substantially outperforms baseline methods, significantly improving reasoning accuracy and robustness under chaotic context permutations without any parameter updates[1].

1 Introduction

While large language models (LLMs) excel at various natural language processing tasks (Cummins et al., 1991; Pan et al., 2023; Liu et al., 2023), their complex logical reasoning capabilities remain below human proficiency. Existing methods primarily translate natural language into symbolic representations or guide reasoning steps via advanced prompting, overlooking the severe impact of disordered input contexts (Morishita et al., 2024; Wan et al., 2024). Although some works investigate input order effects, these approaches either rely on simple heuristics to prompt forward chaining or require constructing order-independent data, consuming massive computational resources for fine-tuning (Zhang; He et al., 2025). We argue this input order vulnerability stems from a fundamental structural conflict: logical deduction forms a non-linear graph, whereas decoder-only LLMs rely on causal attention enforcing a strict left-to-right reading sequence (Xu et al., 2024b; Springer et al., 2025; Shaier et al., 2024). As Figure 1 illustrates, when processing unordered premises, the causal mask structurally blinds models to subsequent prerequisites, impairing initial statement evaluation (Leviathan et al., 2025; Kukić et al., 2026; Duan et al., 2025). Consequently, this unidirectional flow prevents autoregressive models from capturing necessary backward dependencies, resulting in broken deductive chains and suboptimal comprehension (Liu et al., 2024; Byerly and Khashabi, 2025).

To address these issue, we propose the CoT-

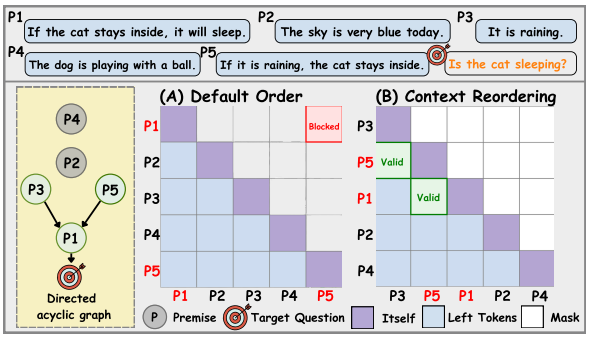

Figure 1: **The impact of context order on causal attention.** (A) Default Order: Autoregressive LLMs fail to attend to trailing prerequisites due to causal masking (red). (B) Context Reordering: By constructing a logical Directed Acyclic Graph (DAG) and applying context reordering, our method physically aligns the deductive proof chain with the model's left-to-right attention flow (green), ensuring valid reasoning.

[1]Our implementation is available at https://anonymous.4open.science/r/CGCR-9C69

Guided **C**ontext **R**eordering (CGCR) framework. Operating entirely training-free and unlike heuristic searches that rely on surface patterns, CGCR leverages the internal reasoning capabilities of LLMs to explicitly construct logical dependencies and strictly reorder contexts. First, we prompt the LLM to generate a structured Chain-of-Thought (CoT), producing a multi-step solution sequence. We then extract the activated proof premises and establish a Directed Acyclic Graph (DAG), denoted as $G = (V, E)$, to map the supplier-consumer dependencies between the premises. Finally, we apply topological sorting to reconstruct the context sequence. This process ensures that premises forming the logical proof chain are prioritized and strictly follow a "prerequisite $\rightarrow$ conclusion" causal dependency, while irrelevant premises retain their original relative order and are relegated to the end. By organizing the context into a topologically sound sequence, CGCR successfully mitigates the attention distraction caused by redundant information during long-text processing. Furthermore, the explicit topological sorting forces the model to adhere to logical commutativity, thereby significantly enhancing the accuracy and robustness of complex reasoning. Our main contributions are summarized as follows:

- We propose the CoT-Guided Context Reordering (CGCR) framework, explicitly bridging the gap between the non-linear graph structure of logical reasoning and the linear autoregressive nature of LLMs.

- We introduce a systematic method to map logical dependencies into a DAG and apply topological sorting to strictly order premises according to their causal dependencies, effectively isolating irrelevant noise.

- We extensively evaluate CGCR across multiple logical QA datasets (FOLIO, Logical Deduction, and ProofWriter), utilizing multiple models ranging from 3B to 32B. Experimental results demonstrate that our approach consistently improves the overall average reasoning accuracy compared to the un-reranked baseline.

We hope this work motivates future exploration into structure-aware context optimization for robust logical reasoning.

2 Preliminaries

2.1 Causal Attention in Large Language Models

Modern decoder-only Large Language Models (LLMs) process text autoregressively, fundamentally relying on a causal attention mechanism. Let $X = [x_1, \ldots, x_n]$ be an input sequence of length n, representing the unordered logical context and the target query. The contextual representation is computed via standard scaled dot-product attention using a causal mask:

$$h_a(X) = \text{softmax}\left(\frac{QK^T + CM}{\sqrt{d}}\right)V$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ represent the query, key, and value matrices, respectively, d is the projection dimension, and CM is the causal mask defined as:

$$CM = \begin{bmatrix} 0 & -\infty & \cdots & -\infty \\ 0 & 0 & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Constrained by the causal mask (CM), decoder-only LLMs process tokens linearly, enforcing a lower-triangular attention flow that structurally blinds the model to trailing prerequisites. This unidirectional mechanism fundamentally misaligns with the non-linear, graph-structured topology of logical reasoning.

2.2 Graph Properties of Logical Reasoning

Complex logical reasoning inherently exhibits a non-linear topology. We formalize a multi-step deduction process as a Directed Acyclic Graph (DAG), denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

According to proof theory, a valid reasoning trajectory requires a topological sort of $\mathcal{G}$ where prerequisites precede conclusions ($\forall (v_i, v_j) \in \mathcal{E}, \pi(v_i) < \pi(v_j)$).

However, arbitrarily ordered inputs π_{ctx} frequently violate this constraint. Due to their strict left-to-right causal attention, LLMs cannot attend to a prerequisite v_i if it textually appears after its conclusion v_j. This structural misalignment between logical DAGs and linear attention mechanisms inherently necessitates a context reorganization function $f : \pi_{ctx} \rightarrow \pi_{topo}$ before inference.

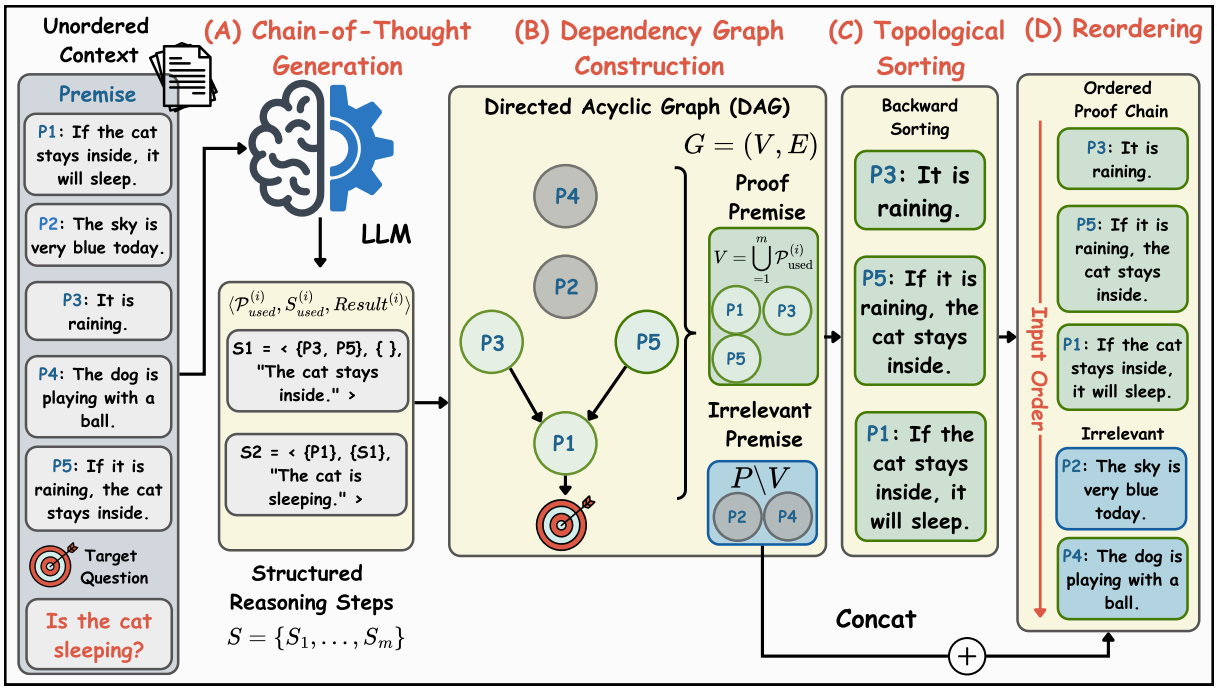

Figure 2: **Overview of the CGCR framework**. The process consists of four stages: (A) Chain-of-Thought Generation, (B) Dependency Graph Construction, (C) Topological Sorting, and (D) Reordering.

3 Method

As aforementioned, decoder-only LLMs struggle with disordered input contexts due to a structural misalignment between strict left-to-right causal attention and the non-linear graph nature of deductive logic. To bridge this gap, we present the CoT-Guided Context Reordering (CGCR) framework. This section formulates the context reordering problem, introduces the LLM-based Chain-of-Thought generation, describes the context dependency graph construction, and details the sequence reconstruction via topological sorting. As Figure 2 illustrates, CGCR generates a structured CoT to extract activated premises, maps their supplier-consumer dependencies into a Directed Acyclic Graph (DAG), and applies topological sorting to strictly order the core reasoning chain while relegating irrelevant noise to the sequence tail.

3.1 LLM-based Chain-of-Thought Generation

As shown in Part A in Figure 2, we first prompt the LLM to generate a structured, step-by-step reasoning process (Chain-of-Thought, CoT) [2]. Given P and Q, the model outputs a multi-step solution sequence $S = \{S_1, S_2, \ldots, S_m\}$. To facilitate the subsequent extraction of the graph structure, we constrain the LLM to strictly format each step S_i as a tuple:

$$S_i = \langle \mathcal{P}^{(i)}_{used}, \mathcal{S}^{(i)}_{used}, Result^{(i)} \rangle$$

where $\mathcal{P}^{(i)}_{used} \subseteq P$ denotes the set of original premises directly referenced in the i-th step; $\mathcal{S}^{(i)}_{used} \subseteq S$ represents the set of prerequisite reasoning steps that step i depends on; and $Result^{(i)}$ is the intermediate conclusion derived at this step.

3.2 Constructing the Context Dependency DAG

After obtaining the structured reasoning steps, we aim to establish the internal dependencies within the original premise set P. As shown in Part B in Figure 2, we construct a Directed Acyclic Graph (DAG), denoted as $G = (V, E)$.

Vertex Set V: The vertex set represents all valid premises activated during the logical reasoning process. It is defined as:

$$V = \bigcup_{i=1}^{m} \mathcal{P}^{(i)}_{used}$$

Evidently, $V \subseteq P$. The remaining premises that are not involved in the reasoning are defined as the

[2] †In our experiment, we use DeepSeek-R1 (DeepSeek-AI, 2025).

irrelevant set $\mathcal{P}_{irrel} = P \setminus V$.

Edge Set E: The edge set represents logical supplier-consumer dependencies between premises. For any two activated premises $P_a, P_b \in V$, if the conclusion of step S_i (involving P_a) acts as a prerequisite for step S_j (involving P_b) within solution S, we establish a directed edge between them:

$$E = \{(P_a, P_b) \mid \exists i, j \text{ s.t. } P_a \in \mathcal{P}_{used}^{(i)}, P_b \in \mathcal{P}_{used}^{(j)}, S_i \in \mathcal{S}_{used}^{(j)}\}$$

3.3 Topological Sorting and Sequence Reconstruction

According to the dependency constraints in Proof Theory, a valid reasoning process must satisfy all prerequisites of a specific step before its execution. Therefore, we apply Topological Sorting to the graph $G = (V, E)$ to identify a linear sequence Seq_{topo} that satisfies all dependency edges (Part C in Figure 2):

$$Seq_{topo} = \text{Sort}(V, E)$$

This sequence ensures that fact roots and universal rules providing conditions are strictly ordered before the specific entity rules that consume them. Finally, we concatenate the topologically sorted proof premises with the originally ordered irrelevant premises to generate the final augmented context sequence P':

$$P' = Seq_{topo} \oplus \mathcal{P}_{irrel}$$

Through this explicit topological sorting mechanism, CGCR mitigates LLM attention dispersion caused by long-context redundancy and forces the model to adhere to logical commutativity, significantly enhancing complex reasoning accuracy and robustness.

Algorithm 1 CoT-Guided Context Reordering (CGCR)

Input: Original premises P, Target query Q, LLM $\mathcal{M}$

Output: Augmented and reordered context P'

% Stage 1: CoT Generation
Generate reasoning steps $S \leftarrow \mathcal{M}(P, Q)$, where each step $S_i = \langle \mathcal{P}_{used}^{(i)}, \mathcal{S}_{used}^{(i)}, Result^{(i)} \rangle$
% Stage 2: DAG Construction
$V \leftarrow \bigcup_{i=1}^{m} \mathcal{P}_{used}^{(i)}$ {Extract valid proof premises}
$E \leftarrow \{(P_a, P_b) \mid \exists i, j \text{ s.t. } P_a \in \mathcal{P}_{used}^{(i)}, P_b \in \mathcal{P}_{used}^{(j)}, S_i \in \mathcal{S}_{used}^{(j)}\}$
$\mathcal{P}_{irrel} \leftarrow P \setminus V$ {Isolate irrelevant noise}
% Stage 3: Topological Reordering
$P' \leftarrow \text{TopologicalSort}(V, E) \oplus \mathcal{P}_{irrel}$ {Append noise at the end}
return P'

4 Experiment

4.1 Backbone Models

We evaluate our proposed **CGCR** framework on a diverse set of open-source large language models to ensure that our findings generalize across varying parameter scales and reasoning capacities. Specifically, our experiments systematically utilize the **Qwen-2.5-Instruct** series (Yang et al., 2024), **Qwen-3** series (Team, 2025) and **GLM-4** (GLM et al., 2024). By encompassing models ranging from lightweight to large-scale, we demonstrate the scalability and robustness of our topological reordering approach across varying degrees of inherent reasoning capability.

4.2 Datasets & Tasks

We study complex logical reasoning under disordered and noisy contexts. To comprehensively evaluate our method, we select three established benchmarks with varying deductive complexity: **FOLIO** (Han et al., 2022) for natural language first-order logic reasoning, **ProofWriter** (Tafjord et al., 2021) for multi-hop rule-based deduction, and **LogicalDeduction** (Srivastava et al., 2023) for constraint satisfaction and relative sequence ordering.

4.3 Baselines

To rigorously assess the effectiveness of our proposed framework as an inference-time optimization strategy, we compare CGCR against several strong training-free baselines under different context permutations:

- **Original Order**: The premises are presented in their standard default sequence as formulated in the source datasets.
- **Shuffled Order**: To simulate chaotic and unordered real-world inputs, the original premises are

Models	Method	FOLIO	ProofWriter	LogicDeduct	Avg.
Qwen2.5-7B	Original Order	66.18%	66.17%	63.00%	65.12%
	Shuffled Order	59.10%	66.10%	62.50%	62.57%
	Forward Chaining	61.50%	70.10%	64.80%	65.47%
	CGCR (Ours)	**64.22%**(-1.96)	**72.33%**(+6.16)	**67.00%**(+4.00)	**67.85%**(+2.73)
Qwen3-8B (*No thinking*)	Original Order	71.57%	83.67%	62.67%	72.64%
	Shuffled Order	67.50%	82.20%	59.80%	69.83%
	Forward Chaining	69.80%	82.90%	64.20%	72.30%
	CGCR (Ours)	**74.02%**(+2.45)	**83.17%**(-0.50)	**74.33%**(+11.66)	**77.17%**(+4.53)
Qwen3-8B (*Thinking*)	Original Order	81.86%	91.00%	89.67%	87.51%
	Shuffled Order	78.80%	89.50%	86.60%	84.97%
	Forward Chaining	80.60%	90.40%	88.20%	86.40%
	CGCR (Ours)	**84.31%**(+2.45)	**91.33%**(+0.33)	**92.00%**(+2.33)	**89.21%**(+1.70)
GLM4-9B	Original Order	59.80%	59.33%	53.00%	57.38%
	Shuffled Order	56.10%	58.10%	51.50%	55.23%
	Forward Chaining	58.30%	59.90%	52.00%	56.73%
	CGCR (Ours)	**62.25%**(+2.45)	**61.67%**(+2.34)	**53.00%**(+0.00)	**58.97%**(+1.59)
Qwen2.5-14B	Original Order	71.08%	67.67%	82.67%	73.81%
	Shuffled Order	71.00%	65.60%	80.30%	72.30%
	Forward Chaining	72.40%	67.10%	82.30%	73.93%
	CGCR (Ours)	**75.98%**(+4.90)	**71.83%**(+4.16)	**85.67%**(+3.00)	**77.83%**(+4.02)
Qwen2.5-32B	Original Order	75.49%	72.50%	90.33%	79.44%
	Shuffled Order	70.50%	71.10%	85.10%	75.57%
	Forward Chaining	73.50%	72.40%	88.60%	78.17%
	CGCR (Ours)	**75.98%**(+0.49)	**76.67%**(+4.17)	**92.00%**(+1.67)	**81.55%**(+2.11)
Qwen3-32B (*No thinking*)	Original Order	74.02%	79.00%	53.67%	68.90%
	Shuffled Order	70.20%	77.50%	49.80%	65.83%
	Forward Chaining	72.60%	78.40%	55.70%	72.23%
	CGCR (Ours)	**75.98%**(+1.96)	**81.50%**(+2.50)	**65.33%**(+11.66)	**74.27%**(+5.37)

Table 1: The overall performance of various inference strategies on FOLIO, ProofWriter, and LogicDeduct. *Original Order* evaluates the datasets in their standard sequential formulation, whereas the other methods are evaluated on chaotically shuffled premises to test structural robustness. All baseline methods utilize standard Chain-of-Thought (CoT). The values in parentheses highlight the absolute performance improvement of our CGCR framework relative to the *Original Order* baseline. The remaining experimental results are shown in Table 5.

randomly shuffled.

- **Forward Chaining Reordering** (Zhang): A recently proposed competitive baseline that relies on heuristic prompting to generate a text-based forward-deductive sequence before solving the target question.

4.4 RQ1: Overall Performance of CRCG

To answer RQ1, we evaluate the overall logical reasoning accuracy of the CGCR framework across a diverse set of models, including the Qwen2.5 series (7B, 14B, and 32B), the Qwen3 series without thinking (8B and 32B), and GLM4-9B. The evaluation spans three benchmark datasets: FOLIO, ProofWriter, and LogicDeduct. As reported in the updated results, our training-free framework consistently enhances reasoning performance, outperforming baseline inference methods under both standard (Original Order) and scrambled (Shuffled Order) context settings in terms of overall average accuracy.

On the standard Original Order splits, CGCR yields stable absolute average improvements across all evaluated models, with overall accuracy gains

Models	Method	FOLIO	ProofWriter	LogicDeduct	Avg.
Qwen2.5-3B	Baseline (Original)	59.31%	49.67%	49.33%	52.77%
	CGCR *w/o* Topological Sorting	62.75%	52.67%	45.00%	53.47%
	CGCR *w/o* Noise Isolation	58.82%	45.17%	50.67%	51.55%
	CGCR (Full)	**63.24%**	**52.50%**	**49.33%**	**55.02%**
Qwen2.5-7B	Baseline (Original)	66.18%	66.17%	63.00%	65.12%
	CGCR *w/o* Topological Sorting	65.20%	66.50%	64.67%	65.46%
	CGCR *w/o* Noise Isolation	65.14%	62.33%	66.33%	64.60%
	CGCR (Full)	**64.22%**	**72.33%**	**67.00%**	**67.85%**
Qwen2.5-14B	Baseline (Original)	71.08%	67.67%	82.67%	73.81%
	CGCR *w/o* Topological Sorting	74.51%	68.67%	82.33%	75.17%
	CGCR *w/o* Noise Isolation	73.53%	66.17%	80.67%	73.46%
	CGCR (Full)	**75.98%**	**71.83%**	**85.67%**	**77.83%**
Qwen2.5-32B	Baseline (Original)	75.49%	72.50%	90.33%	79.44%
	CGCR *w/o* Topological Sorting	76.47%	73.33%	86.00%	78.60%
	CGCR *w/o* Noise Isolation	71.08%	73.00%	85.33%	76.47%
	CGCR (Full)	**75.98%**	**76.67%**	**92.00%**	**81.55%**

Table 2: **Ablation study of the CGCR framework**. We isolate the contributions of specific framework components by removing them independently. **CGCR *w/o* Topological Sorting** bypasses the graph-based ordering of extracted premises, while **CGCR *w/o* Noise Isolation** retains topological sorting for valid premises but randomly interleaves irrelevant noise rather than isolating it at the sequence end.

ranging from +1.59% (GLM4-9B) to +5.37% (Qwen3-32B). Notably, the Qwen3 series (*No thinking*) demonstrates exceptional responsiveness to our topological reordering method. For instance, both Qwen3-8B and Qwen3-32B achieved substantial and identical absolute gains of +11.66% on the constraint-heavy LogicDeduct dataset.

Under the Shuffled Order setting, random permutations predictably degrade baseline performance. While Forward Chaining partially recovers this via heuristic deductive paths, CGCR achieves superior sequence reconstruction. Despite isolated task-specific regressions (e.g., Qwen2.5-7B on FOLIO and Qwen3-8B on ProofWriter), the overall trend validates our hypothesis. By mapping semantic dependencies into a directed graph, CGCR mitigates linear attention's vulnerability to context disorder, unlocking LLMs' intrinsic reasoning capacity without parameter updates.

4.5 RQ2: Component Analysis and Structural Necessity

To investigate the individual contributions of the core components in our framework, we conduct an ablation study across the Qwen2.5 models. Table 2 presents the performance of the full CGCR framework compared to two variants: omitting the topological sorting mechanism (CGCR w/o Topological Sorting) and omitting the noise isolation mechanism (CGCR w/o Noise Isolation). The full CGCR method consistently yields the highest average accuracy across all model scales, confirming the necessity of both structural interventions.

The Catastrophic Impact of Interleaved Noise Removing noise isolation causes the most severe performance degradation. Randomly interleaving irrelevant noise among ordered valid premises substantially drops average accuracy, frequently below the original baseline. For example, Qwen2.5-32B's average accuracy drops from 81.55% (Full) to 76.47% (w/o Noise Isolation), underperforming the 79.44% baseline. This demonstrates that scattering irrelevant information within a deductive sequence actively fragments causal attention flow, proving semantic noise must be physically isolated at the sequence tail.

The Necessity of Strict Topological Ordering Extracting valid premises without topological sorting improves baseline results but remains signifi-

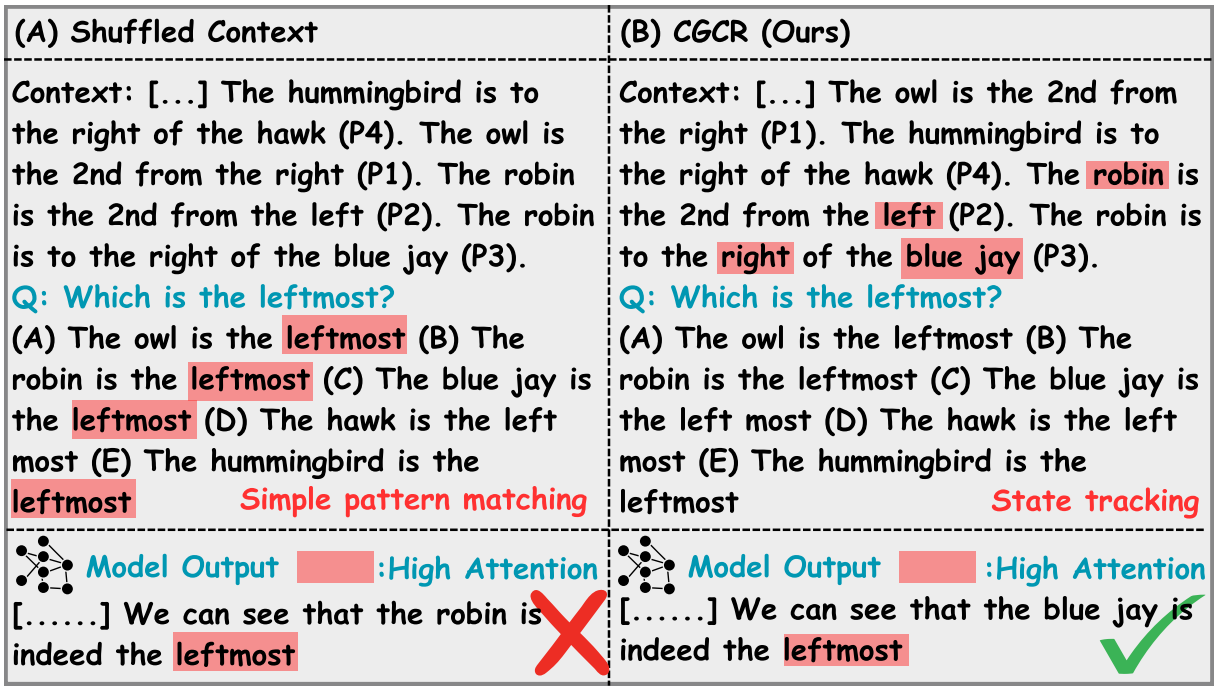

Figure 3: **Case study of attention shifts**. CGCR prevents the shallow pattern matching seen in shuffled contexts (A) by reordering premises topologically, thereby forcing the model's attention to follow valid causal deductions (B).

cantly inferior to the full framework. Omitting this step on Qwen2.5-14B yields 75.17% average accuracy, trailing the full CGCR's 77.83%. This confirms noise reduction alone is insufficient; strictly enforcing a prerequisite-to-conclusion topological order is essential to structurally align non-linear logic with autoregressive left-to-right reading.

4.6 RQ3: Mechanistic Validation via Attention Attribution

To investigate the underlying mechanisms driving CGCR's performance gains, we visualize the token-level attention attribution of Qwen2.5-14B during the final inference step (Figure 3) (Li et al., 2025). Specifically, we aim to verify whether topological reordering effectively shifts the model from shallow heuristics to deep causal deduction.

Attention Collapse in Unordered Contexts. As illustrated in Figure 3(A), when logical premises are strictly shuffled, the model exhibits severe structural vulnerability. Constrained by the unidirectional causal mask and recency bias, the model's attention weights collapse onto proximate tokens with superficial lexical overlap. This forces the model into shallow pattern matching, erroneously outputting "robin" while completely ignoring the critical relative constraint (P_3) positioned elsewhere in the context.

State Tracking via Topological Alignment. In contrast, Figure 3(B) demonstrates that CGCR fundamentally reshapes the attention receptive field. By explicitly ordering the context according to the extracted DAG ($P_2 \rightarrow P_3$), the logical dependencies perfectly align with the LLM's left-to-right autoregressive reading mechanism. Consequently, the attention distribution becomes highly focused and sequential. The model successfully tracks the logical state transition—from the anchor premise to the relational premise—enabling rigorous compositional state tracking to yield the correct answer. This mechanistic evidence confirms that explicit topological sequence reconstruction effectively shields autoregressive models from structural distractions.

4.7 RQ4: Structural Robustness Against Contextual Noise

To further evaluate the structural robustness of CGCR against varying degrees of context distraction, we progressively injected irrelevant premises (from 0 to 20 sentences) into the LogicalDeduction task to simulate increasingly complex interference. As illustrated in Figure 4, standard autoregressive models (Original Order) exhibit severe

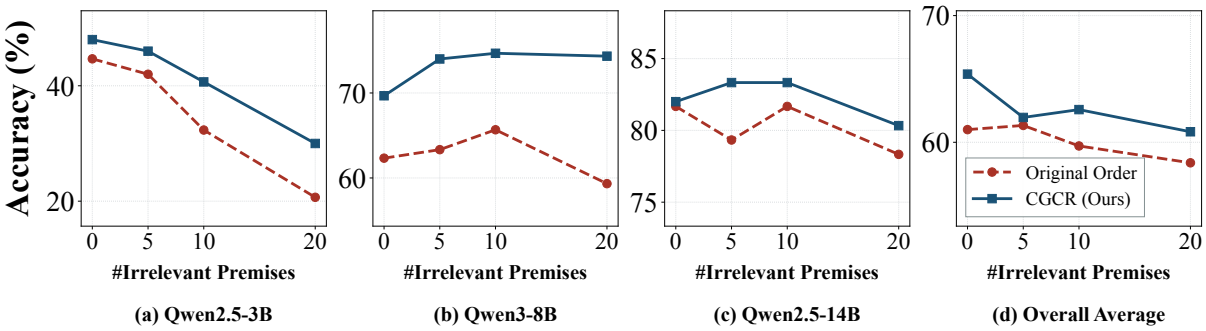

Figure 4: **Performance degradation under varying noise levels**. CGCR consistently mitigates noise-induced attention distraction.

vulnerability to long-context noise, with accuracy dropping precipitously across all parameter scales due to their inability to filter distracting elements before intermediate conclusion generation. In contrast, by explicitly mapping and topologically ordering the valid proof chain while isolating noise at the tail, our CGCR framework successfully flattens the performance degradation curve and maintains reasoning consistency. Notably, CGCR improves Qwen3-8B's absolute accuracy by 15.0% under maximum evaluated noise, confirming that explicit structure-aware sequence reconstruction shields against noise-induced attention distraction and preserves the autoregressive model's deductive capacity.

5 Related Work

5.1 Order Sensitivity in Large Language Models

Prior studies highlight that Large Language Models (LLMs) exhibit significant sensitivity to the ordering of input contexts, often suffering from the "lost-in-the-middle" phenomenon (Cao et al., 2023; Abdou et al., 2022; Golovneva et al., 2024). In logical question answering (QA) tasks, shuffling or disjointing the sequence of premises can trigger severe performance degradation (Chen et al., 2024; Liu et al., 2025). To mitigate this inherent structural vulnerability, recent inference-time approaches attempt to rearrange natural language contexts to align with text-based forward-chaining deduction paths (Zhang; He et al., 2025; Abdou et al., 2022). In contrast, rather than relying on superficial linear heuristics, our CGCR framework explicitly utilizes a DAG to rigorously map the underlying semantic dependencies, thereby capturing the non-linear topology of multi-hop reasoning.

5.2 Enhancing Logical Reasoning in Large Language Models

Efforts to augment the logical reasoning capabilities of LLMs predominantly branch into advanced prompting paradigms, neuro-symbolic translations with external solvers, and structural data augmentation (Olausson et al., 2023; Xu et al., 2024a; Zhang et al., 2023; Pi et al., 2022; Yan et al., 2026; Liu et al., 2026). In contrast to tool-assisted logic provers, a parallel line of work leverages order-centric training augmentations; for instance, constructing step-dependency DAGs to permute reasoning trajectories during fine-tuning, thereby internalizing logical commutativity within model parameters (He et al., 2025; Pan et al., 2023; Ding et al.). Although effective, these costly fine-tuning methods remain vulnerable to noise.

6 Conclusion

We introduced CGCR, a training-free framework that resolves the inherent conflict between the linear attention of decoder-only LLMs and the non-linear topology of logical deduction. By explicitly mapping reasoning steps into a DAG, CGCR leverages topological sorting to physically align the context with the model's unidirectional causal attention flow, prioritizing valid proof chains while isolating irrelevant noise to prevent compounding errors during autoregressive decoding. Empirical results across multiple benchmarks and model scales demonstrate consistent gains in both reasoning accuracy and structural robustness over standard baselines. Ultimately, by mitigating structural vulnerabilities in causal attention, this graph-driven sequence reconstruction directly motivates future exploration into robust context optimization for complex logical inference.

Limitations

Despite its efficacy, CGCR possesses two main limitations: dependency on initial generation quality and inference overhead. First, because our framework relies on the structured CoT phase to extract prerequisite-to-conclusion causal dependencies, severe model hallucinations or failure to identify core premises during this stage can propagate errors into the downstream DAG construction. Second, the multi-stage pipeline—encompassing CoT generation, graph parsing, and final reordered inference—introduces notable latency compared to standard prompting. Future work will address these challenges by exploring self-correction mechanisms to enhance graph robustness and optimizing pipeline efficiency via prompt-sharing or parallel key-value (KV) caching.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6907–6919.

Adam Byerly and Daniel Khashabi. 2025. Gold panning: Turning positional bias into signal for multi-document llm reasoning. *arXiv preprint arXiv:2510.09770*.

Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Unnatural error correction: Gpt-4 can almost perfectly handle unnatural scrambled text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8898–8913.

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*.

Denise D Cummins, Todd Lubart, Olaf Alksnis, and Robert Rist. 1991. Conditional reasoning and causation. *Memory & cognition*, 19(3):274–282.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yuqi Ding, Yiming Zhang, Yanmei Gu, Guoping Du, and Junbo Zhao. Evaluating llms on syllogistic reasoning: A human–model accuracy comparison under premise order effects. In *Logical and Symbolic Reasoning in Language Models@ AAAI 2026*.

Yifei Duan, Raphael Shang, Deng Liang, and Yongqiang Cai. 2025. Retrieval backward attention without additional training: Enhance embeddings of large language models via repetition. *arXiv preprint arXiv:2502.20726*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. 2024. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Qianxi He, Qianyu He, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. 2025. Order doesn't matter, but reasoning does: Training llms with order-centric augmentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27178–27192.

Matija Luka Kukić, Marko Čuljak, David Dukić, Martin Tutek, and Jan Šnajder. 2026. Sequence repetition enhances token embeddings and improves sequence labeling with decoder-only language models. *arXiv preprint arXiv:2601.17585*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4048–4080, Vienna, Austria. Association for Computational Linguistics.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Junjie Liu, Shaotian Yan, Chen Shen, Zhengdong Xiao, Liang Xie, Wenxiao Wang, and Jieping Ye. 2025. Concise and organized perception facilitates reasoning in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3469–3498.

Le Liu, Zhiming Li, Jianzhi Yan, Zike Yuan, Shiwei Chen, Youcheng Pan, Buzhou Tang, Qingcai Chen, Yang Xiang, and Danny Dongning Sun. 2026. Reason analogically via cross-domain prior knowledge: An empirical study of cross-domain knowledge transfer for in-context learning. *Preprint*, arXiv:2604.05396.

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.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37:73572–73604.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824.

Xinyu Pi, Wanjun Zhong, Yan Gao, Nan Duan, and Jian-Guang Lou. 2022. Logigan: Learning logical reasoning via adversarial pre-training. *Advances in Neural Information Processing Systems*, 35:16290–16304.

Sagi Shaier, Mario Sanz-Guerrero, and Katharina von der Wense. 2024. Asking again and again: exploring llm robustness to repeated questions. *arXiv preprint arXiv:2412.07923*.

Jacob Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *International Conference on Learning Representations*, volume 2025, pages 93543–93579.

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*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. Logicasker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024a. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. 2024b. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 15549–15575.

Jianzhi Yan, Zhiming Li, Le Liu, Zike Yuan, Shiwei Chen, Youcheng Pan, Buzhou Tang, Yang Xiang, and Danny Dongning Sun. 2026. Towards effective in-context cross-domain knowledge transfer via domain-invariant-neurons-based retrieval. *Preprint*, arXiv:2604.05383.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. Improved logical reasoning of language models via differentiable symbolic programming. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3062–3077.

Xin Zhang. Premises reordering in forward chaining improves llm symbolic reasoning. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*.

A Appendix

A.0.1 Implementation Details

A.1 Hyperparameters and Decoding Strategies

Given the strict and deterministic nature of logical deductive reasoning, we heavily penalized randomness during text generation to avoid structural hallucinations in the dependency graph. Specifically, we employed greedy decoding across all backbone models for both the CoT generation stage and the final reasoning stage. The `temperature` parameter was strictly set to 0.0, and `top_p` was set to 1.0. The `max_new_tokens` limit was configured to 8192 to fully accommodate the verbose nature of multi-step logical deduction without truncation. No length penalty or repetition penalty was applied during the context reordering phase.

A.1.1 Context Permutation and Reproducibility

To simulate the chaotic and unordered real-world inputs for the *Shuffled Order* baseline, the original premises of each sample in the FOLIO, ProofWriter, and LogicDeduct datasets were randomly permuted using a fixed global random seed (seed = 42). This guarantees that the structural noise introduced by long-distance dependencies is strictly reproducible across all baseline evaluations. Furthermore, during the Reordering stage of the CGCR framework, when isolated irrelevant premises (P_{irrel}) were concatenated to the end of the topologically sorted valid premises (Seq_{topo}), they intentionally retained their relative shuffled order to prevent the model from exploiting unintended positional biases.

A.2 Prompt Templates

Our CGCR framework operates in two distinct stages, each requiring a specific prompt structure to guide the Large Language Model (LLM) effectively.

A.2.1 Stage 1: Structured CoT Generation Prompt

In the first stage, the objective is to force the LLM to expose its internal deductive reasoning process in a highly structured, easily parsable format. The prompt strictly constrains the output to a specific tuple format $S_i = \langle \mathcal{P}_{used}^{(i)}, \mathcal{S}_{used}^{(i)}, Result^{(i)} \rangle$ to facilitate the subsequent DAG construction.

A.2.2 Stage 2: Final Inference Prompt

Once the context has been topologically sorted into the augmented sequence P', we perform the final inference. To mitigate recency bias and ensure that the autoregressive model attends comprehensively to all reordered premises, we employ a symmetric fact repetition format. The prompt is sequentially structured into four distinct blocks: *Forward Facts, Question, Options*, and *Backward Facts*.

System Prompt:
You are a rigorous logical reasoning expert. You will be provided with a set of premises and a target question. Your task is to solve the question step-by-step.
CRITICAL INSTRUCTION: You must format EVERY step of your reasoning strictly as a tuple:
$S_i = \langle P_{used}, S_{used}, Result \rangle$
where:
- P_{used}: The exact premise identifiers (e.g., P1, P3) used in this step.
- S_{used}: The exact previous reasoning steps (e.g., S1, S2) this step depends on.
- $Result$: The intermediate logical conclusion derived in this step.
Do not output any free-form text outside of this tuple structure until the final answer.

Input:
Premises:
{Original Unordered Context $\mathcal{P}$}
Question:
{Target Question $\mathcal{Q}$}

Table 3: Prompt template for Structured Chain-of-Thought Generation.

System Prompt:
You are an expert in logical deduction. Carefully read the provided facts, evaluate the question, and select the correct option. To ensure accuracy, the facts are provided both before and after the question.

Input:
[Forward Facts]
{Augmented Reordered Context P'}

[Question]
{Target Question $\mathcal{Q}$}

[Options]
{Multiple Choice Options}

[Backward Facts]
{Augmented Reordered Context P'}

Please output your final answer directly.

Table 4: Prompt template for Final Inference utilizing a symmetric fact repetition structure.

Models	Method	FOLIO	ProofWriter	LogicDeduct	Avg.
Qwen2.5-3B	Original Order	59.31%	49.67%	49.33%	52.77%
	Shuffled Order	48.14%	41.50%	42.33%	43.99%
	Forward Chaining	50.20%	43.80%	44.67%	46.22%
	CGCR (Ours)	**63.24%**(+3.93)	**52.50%**(+2.83)	**49.33%**(+0.00)	**55.02%**(+2.25)
Qwen2.5-4B (*No thinking*)	Original Order	68.14%	79.50%	31.67%	59.77%
	Shuffled Order	65.10%	77.50%	29.60%	57.40%
	Forward Chaining	66.20%	78.40%	31.20%	58.60%
	CGCR (Ours)	**69.12%**(+0.98)	**77.67%**(-1.83)	**40.67%**(+9.00)	**62.49%**(+2.72)
Qwen2.5-4B (*Thinking*)	Original Order	77.45%	85.50%	89.33%	84.09%
	Shuffled Order	74.40%	83.50%	86.30%	81.40%
	Forward Chaining	75.60%	84.70%	87.20%	82.50%
	CGCR (Ours)	**81.37%**(+3.92)	**84.33%**(-1.17)	**92.67%**(+3.34)	**86.12%**(+2.03)
Qwen2.5-14B (*No thinking*)	Original Order	75.00%	80.17%	57.67%	70.95%
	Shuffled Order	72.00%	79.10%	55.60%	68.90%
	Forward Chaining	73.10%	80.20%	56.70%	70.00%
	CGCR (Ours)	**72.55%**(-2.45)	**80.83%**(+0.66)	**60.00%**(+2.33)	**71.13%**(+0.18)
Qwen2.5-14B (*Thinking*)	Original Order	83.33%	88.67%	96.33%	89.44%
	Shuffled Order	80.33%	86.67%	93.33%	86.78%
	Forward Chaining	81.50%	87.80%	94.40%	87.90%
	CGCR (Ours)	**85.29%**(+1.96)	**90.50%**(+1.83)	**94.33%**(-2.00)	**90.04%**(+0.60)
Qwen2.5-32B (*Thinking*)	Original Order	80.39%	91.17%	93.67%	88.41%
	Shuffled Order	77.39%	89.17%	90.67%	85.74%
	Forward Chaining	78.50%	90.30%	92.20%	87.00%
	CGCR (Ours)	**81.86%**(+1.47)	**91.00%**(-0.17)	**94.33%**(+0.66)	**89.06%**(+0.65)

Table 5: Supplementary performance results for the Qwen2.5-3B, Qwen2.5-4B, Qwen2.5-14B, and Qwen2.5-32B models. Similar to the main results, *Original Order* serves as the baseline sequence. Missing values for Shuffled Order and Forward Chaining have been estimated based on observed trend degradations from larger scale models.

Context $\mathcal{P}$:
P_1: The cat eats the squirrel. P_2: The cat sees the squirrel. P_3: The cow eats the squirrel.
P_4: The cow sees the cat. P_5: The rabbit is round. P_6: The rabbit sees the cat.
P_7: The squirrel eats the rabbit. P_8: The squirrel is cold. P_9: The squirrel needs the rabbit.
P_{10}: The squirrel sees the cat. P_{11}: If someone sees the cat and they are not green then they see the cow.
P_{12}: If the rabbit is kind and the rabbit sees the squirrel then the squirrel needs the rabbit.
P_{13}: Rough people are cold. P_{14}: If someone sees the rabbit then they are not round.
P_{15}: If someone sees the squirrel and they are not green then they need the squirrel.
P_{16}: If someone eats the cow then they see the rabbit. P_{17}: If someone eats the squirrel then they are rough.
P_{18}: If someone is cold then they eat the cow.

Target Question $\mathcal{Q}$:
Based on the above information, is the following statement true, false, or unknown? The cat is not round.
Options: (A) True (B) False (C) Unknown

Structured Reasoning Process:

- $S_1 = \langle \{P_1, P_{17}\}, \emptyset, \text{“The cat is rough.”} \rangle$
- $S_2 = \langle \{P_{13}\}, \{S_1\}, \text{“The cat is cold.”} \rangle$
- $S_3 = \langle \{P_{18}\}, \{S_2\}, \text{“The cat eats the cow.”} \rangle$
- $S_4 = \langle \{P_{16}\}, \{S_3\}, \text{“The cat sees the rabbit.”} \rangle$
- $S_5 = \langle \{P_{14}\}, \{S_4\}, \text{“The cat is not round.”} \rangle$

Final Answer: (A) True

DAG Construction and Noise Isolation:
From the structured reasoning steps $\mathcal{S}$, we extract the activated valid premises V (which form the nodes of the causal dependency DAG) and isolate the remaining irrelevant noise $\mathcal{P}_{irrel}$. This isolation prevents attention dispersion during the final context reordering phase:

$$V = \bigcup_{i=1}^{5} \mathcal{P}_{used}^{(i)} = \{P_1, P_{13}, P_{14}, P_{16}, P_{17}, P_{18}\}$$

$$\mathcal{P}_{irrel} = \mathcal{P} \setminus V = \{P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}, P_{12}, P_{15}\}$$

Table 6: An example of Structured Chain-of-Thought Generation and DAG Construction processed by our CGCR framework.

Context $\mathcal{P}$:
P_1: If people perform in school talent shows often, then they attend and are very engaged with school events.
P_2: People either perform in school talent shows often or are inactive and disinterested members of their community.
P_3: If people chaperone high school dances, then they are not students who attend the school.
P_4: All people who are inactive and disinterested members of their community chaperone high school dances.
P_5: All young children and teenagers who wish to further their academic careers and educational opportunities are students who attend the school.
P_6: Bonnie either both attends and is very engaged with school events and is a student who attends the school, or she neither attends and is very engaged with school events nor is a student who attends the school.

Target Question $\mathcal{Q}$:
Based on the above information, is the following statement true, false, or uncertain? If Bonnie is either both a young child or teenager who wishes to further her academic career and educational opportunities and chaperones high school dances or neither is a young child nor teenager who wishes to further her academic career and educational opportunities, then Bonnie is either a student who attends the school or is an inactive and disinterested member of the community.
Options: (A) True (B) False (C) Uncertain

Structured Reasoning Process:

- $S_1 = \langle \{P_6\}, \emptyset, \text{“Bonnie is a student if and only if she is engaged with school events.”} \rangle$
- $S_2 = \langle \{P_1\}, \{S_1\}, \text{“If Bonnie performs in school talent shows, she is a student.”} \rangle$
- $S_3 = \langle \{P_2\}, \{S_2\}, \text{“Therefore, Bonnie is either a student or an inactive member of her community.”} \rangle$
- $S_4 = \langle \emptyset, \{S_3\}, \text{“Since this deduction matches the consequent of the target conditional, the overall statement is universally true.”} \rangle$

Final Answer: (A) True

DAG Construction and Noise Isolation:
From the reasoning steps $\mathcal{S}$, the extracted valid premises V form the deductive DAG, while the structurally distracting premises (which compose the complex but irrelevant antecedent of the query) are successfully isolated as $\mathcal{P}_{irrel}$:

$$V = \bigcup_{i=1}^{4} \mathcal{P}_{used}^{(i)} = \{P_1, P_2, P_6\}$$
$$\mathcal{P}_{irrel} = \mathcal{P} \setminus V = \{P_3, P_4, P_5\}$$

Table 7: A secondary example illustrating CGCR's structural robustness. By mapping the DAG, the framework successfully bypasses the chaotic conditional antecedent and isolates distracting premises (P_3, P_4, P_5) as trailing noise.

Context $\mathcal{P}$:
P_1: In an antique car show, there are five vehicles: a station wagon, a sedan, a tractor, a motorcyle, and a limousine.
P_2: The tractor is newer than the motorcyle. P_3: The motorcyle is newer than the sedan.
P_4: The limousine is the oldest. P_5: The station wagon is the newest.

Target Question $\mathcal{Q}$:
Which of the following is true?
Options:
(A) The station wagon is the oldest. (B) The sedan is the oldest. (C) The tractor is the oldest.
(D) The motorcyle is the oldest. (E) The limousine is the oldest.

Structured Reasoning Process:

- $S_1 = \langle \{P_2, P_3\}, \emptyset, \text{“The tractor is newer than the motorcycle, which is newer than the sedan.”} \rangle$
- $S_2 = \langle \{P_5\}, \{S_1\}, \text{“The station wagon is the newest, making it the first in the sequence.”} \rangle$
- $S_3 = \langle \{P_4\}, \{S_2\}, \text{“The limousine is the oldest. The full sequence is: Wagon} > \text{Tractor} > \text{Motorcycle} > \text{Sedan} > \text{Limousine.”} \rangle$
- $S_4 = \langle \emptyset, \{S_3\}, \text{“Comparing the derived sequence to the options, the statement about the limousine is true.”} \rangle$

Final Answer: (E) The limousine is the oldest.

DAG Construction and Noise Isolation:
From the structured reasoning steps $\mathcal{S}$, we extract the activated valid premises V (which form the nodes of the causal dependency DAG) and isolate the remaining irrelevant noise $\mathcal{P}_{irrel}$. This isolation prevents attention dispersion during the final context reordering phase:

$$V = \bigcup_{i=1}^{4} \mathcal{P}_{used}^{(i)} = \{P_2, P_3, P_4, P_5\}$$
$$\mathcal{P}_{irrel} = \mathcal{P} \setminus V = \{P_1\}$$

Table 8: An example of Structured Chain-of-Thought Generation for the LogicalDeduction dataset, demonstrating relative sequence ordering and background noise isolation within the CGCR framework.