Bridging Associative Memory and Logical Reasoning: A Causal-Enhanced Dual-Process Approach with Metacognitive Monitoring

Jiali Zhang (zhangjiali2@mails.gdut.edu.cn)
Chong Chen[1] (chenc2021@gdut.edu.cn)
Tao Wang (wangtao_cps@gdut.edu.cn)
Zhuowei Wang (zwwang@gdut.edu.cn)
Lianglun Cheng (LLCheng@gdut.edu.cn)
Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology
Guangzhou 510006, China

Abstract

While Large Language Models (LLMs) exhibit impressive fluency, their reasoning often relies on surface-level statistical correlations rather than robust causal understanding. Current reasoning approaches, such as Retrieval-Augmented Generation (RAG), mitigate knowledge obsolescence but typically depend on vector-based retrieval, which mimics associative memory but fails to support the rigorous logical deduction required for complex queries. To address this, we propose a novel framework titled Bridging Associative Memory and Logical Reasoning, which implements a dual-process cognitive architecture. We construct high-fidelity Causal Knowledge Graphs through a human-in-the-loop method, utilizing expert-supervised LoRA fine-tuning to extract reliable causal dependencies from raw text. During inference, a Query Cognitive Planner orchestrates a Dual-Path Memory Access strategy, synergizing associative vector evidence with causal graph traversal. Crucially, a metacognitive validation acts as a "System 2" critic for iterative self-correction. Experiments demonstrate that this cooperative interaction outperforms associative baselines, confirming that explicit "System 2" monitoring effectively reduces logical hallucinations.

Keywords: Large Language Models; Retrieval-Augmented Generation; Causal Knowledge Graph; Dual-Process Theory; Metacognition

Introduction

Large Language Models (LLMs) have achieved unprecedented success in recent years, while continuing to struggle with rigorous logical reasoning, often relying on probabilistic correlations rather than genuine causal understanding (Chi et al., 2024). This limitation leads to "hallucinations", where models generate plausible but factually incorrect or logically incoherent statements, particularly in high-stakes domains such as healthcare and industrial intelligence (Zhang et al., 2025). Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to mitigate knowledge obsolescence and hallucinations by grounding generation in external corpora (Lewis et al., 2020). However, standard RAG systems predominantly employ dense vector retrieval based on semantic similarity. As noted by recent studies, semantic proximity does not equate to logical entailment (Shao et al., 2023). When grounded in standard RAG, models function akin to a human utilizing "associative memory", effectively retrieving thematically related documents but failing to extract the specific causal mechanisms required for multi-hop reasoning. Consequently, when faced with complex queries, traditional RAG systems often retrieve disjointed evidence, leading to reasoning disconnects or the hallucination of non-existent causal links (Vu et al., 2024).

Cognitive science offers a theoretical blueprint for addressing these limitations through the "Dual-Process Theory" (Karpukhin et al., 2020; Evans & Stanovich, 2013). It posits that human cognition comprises two distinct systems: System 1 (fast, intuitive, and associative) and System 2 (slow, analytical, and logical). Current LLM-based RAG workflows largely mirror System 1, prioritizing rapid retrieval and generation over deliberate verification. Recent neuro-symbolic research suggests that integrating structured logical representations can effectively simulate System 2 reasoning capabilities in neural models (Garcez & Lamb, 2023).

Drawing upon this cognitive architecture, we propose a novel framework titled Bridging Associative Memory and Logical Reasoning. Our approach bridges the gap between statistical probability and logical validity by synergizing unstructured semantic retrieval with structured causal reasoning. We propose a robust reasoning system that necessitates two distinct memory components: a Vector Database acting as associative memory for broad context, and a Causal Knowledge Graph (CKG) serving as long-term logical memory for precise deduction. As illustrated in our technical roadmap, our framework operates in two phases. First, addressing the noise in automated knowledge extraction, we introduce a High-Fidelity Causal Graph Acquisition method. We employ a Human-in-the-Loop strategy to fine-tune a local model via Low-Rank Adaptation (LoRA) (Hu et al., 2022), distilling complex causal pairs from raw texts under expert supervision. Second, during inference, we model implement a Dual-Process Inference mechanism. Unlike standard "retrieve-then-generate" methods, our system utilizes a Query Cognitive Planner to orchestrate a dual-path memory access strategy. Subsequently, a dual-path Metacognitive Validation module acts as a rigorous critic (Du, Li, Torralba, Tenenbaum, & Mordatch, 2023). It decomposes generated answers into atomic claims, verifying them against both factual evidence and causal logic, and triggers a feedback-driven optimization loop to self-correct errors. The main contributions of this study are as follows:

- We propose a unified cognitive framework grounded in Dual-Process Theory, synergizing associative vector mem-

[1]Corresponding author: Chong Chen

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

ory with structured causal graph to enable human-like reasoning for complex queries.

- We develop an expert-supervised method using LoRA fine-tuning to construct reliable Causal Knowledge Graphs, overcoming the noise inherent in zero-shot extraction.
- We introduce a novel dual-path Metacognitive validation loop that explicitly quantifies both causal consistency and factual grounding. This mechanism enables iterative self-correction, significantly reducing logical hallucinations compared to purely associative baselines.

Related Work

External Memory Systems

Standard Retrieval-Augmented Generation (RAG) extends the parametric memory of LLMs by accessing external corpora (Lewis et al., 2020; Guu, Lee, Tung, Pasupat, & Chang, 2020). From a cognitive standpoint, dense vector retrieval functions as an External Associative Memory (Karpukhin et al., 2020). Recent advancements have shifted towards Modular RAG, which incorporates iterative and recursive retrieval strategies to handle complex information needs (Y. Gao et al., 2023; Shao et al., 2023). Despite these improvements, vector-based retrieval struggles with "reasoning across documents" (Xiong et al., 2020). To address this, Graph-based Memory approaches have been proposed to provide structural context. Methods like GraphRAG (Edge et al., 2024), LightRAG (Guo, Xia, Yu, Ao, & Huang, 2024) and HippoRAG2 (Gutiérrez, Shu, Qi, Zhou, & Su, 2025) utilize global graph structures to simulate associative pathways. However, existing Graph-based Memory approaches primarily rely on general-purpose KGs, which capture relational data rather than causal mechanisms. As highlighted by (Z. Jin et al., 2023), reasoning without explicit causal directionality limits the model's ability to answer "why" and "what-if" questions. Our work distinguishes itself by constructing a specialized Causal Knowledge Graph and employing a Dual-Path Memory Access mechanism that mimics the interplay between associative and logical memory.

Causal Reasoning and Extraction

Causal reasoning, the ability to identify mechanisms rather than mere correlations, is fundamental to general intelligence (Z. Jin et al., 2023; Fu, Bao, Zhang, Hu, & Zhang, 2025). While LLMs demonstrate impressive few-shot performance on common-sense reasoning tasks, they frequently falter in rigorous causal deduction (Zečević, Willig, Dhami, & Kersting, 2023). Traditional causal extraction relies on pattern matching or supervised learning, which suffers from limited scalability. Recent works have explored using LLMs for zero-shot causal discovery (Kiciman, Ness, Sharma, & Tan, 2023; Long, Schuster, & Piché, 2023). However, these approaches are prone to significant noise and hallucinations when applied to domain-specific raw texts. For instance, (J. Gao, Ding, Qin, & Liu, 2023) found that LLMs often infer causality from purely temporal sequences. Unlike methods that rely solely on in-context learning, our approach introduces a Human-in-the-Loop paradigm combined with LoRA fine-tuning, ensuring the construction of a high-fidelity causal graph that serves as a reliable logical scaffold for downstream reasoning.

Metacognition and System 2 Reasoning in LLM

Metacognition refers to the ability of a system to monitor and regulate its own cognitive processes. In the context of LLMs, this aligns with System 2 reasoning, characterized by slow, deliberate thought (Bengio, Lecun, & Hinton, 2021). Several frameworks have attempted to induce System 2 behaviors in LLMs. Chain-of-Thought (CoT) (Wei et al., 2022) and Tree of Thoughts (ToT) (Yao et al., 2023) encourage models to generate intermediate reasoning steps. To further enhance reliability, Self-Correction mechanisms have been proposed, acting as primitive metacognitive monitors. Approaches like Reflexion (Shinn, Cassano, Gopinath, Narasimhan, & Yao, 2023), RARR (L. Gao et al., 2023), and CRITIC (Gou et al., 2023) employ external tools or self-feedback to refine outputs. More recently, Self-RAG (Asai, Wu, Wang, Sil, & Hajishirzi, 2024) advanced this paradigm by embedding self-reflection tokens to critique the model's own retrieval quality.

However, most self-correction methods rely on the model's internal confidence or generic scalar feedback, which can be uncalibrated (Huang et al., 2023). They often lack a structured "ground truth" logic to verify against. Our framework formalizes metacognitive monitoring through an explicit Metacognitive Validation module. By calculating granular scores for both causal validity and factual consistency, we implement a computational model of "slow thinking" that specifically targets and corrects the logical disconnects prevalent in "fast thinking" generation.

Method

To address the limitations of standard associative models, which often rely solely on surface-level semantic matching and lack rigorous logical constraints, we propose a framework that bridges associative memory and logical reasoning. Aligning with the "Dual-Process Theory", our approach mimics human cognition by integrating associative memory (System 1, implemented via vector retrieval) with logical reasoning (System 2, implemented via causal graphs and metacognitive monitoring).

As illustrated in Figure 1, this framework solves the problem of logical disconnects and hallucinations in complex reasoning tasks through two distinct yet interconnected phases: (1) Causal Knowledge Graph Acquisition, which builds a reliable structured logical foundation, and (2) Causal-Enhanced Dual-Process Inference, which executes precise reasoning through cognitive planning and a self-correcting metacognitive validation loop.

Causal Knowledge Graph Acquisition

We construct a high-fidelity Causal Knowledge Graph (CKG) combining expert supervision with automated model adapta-

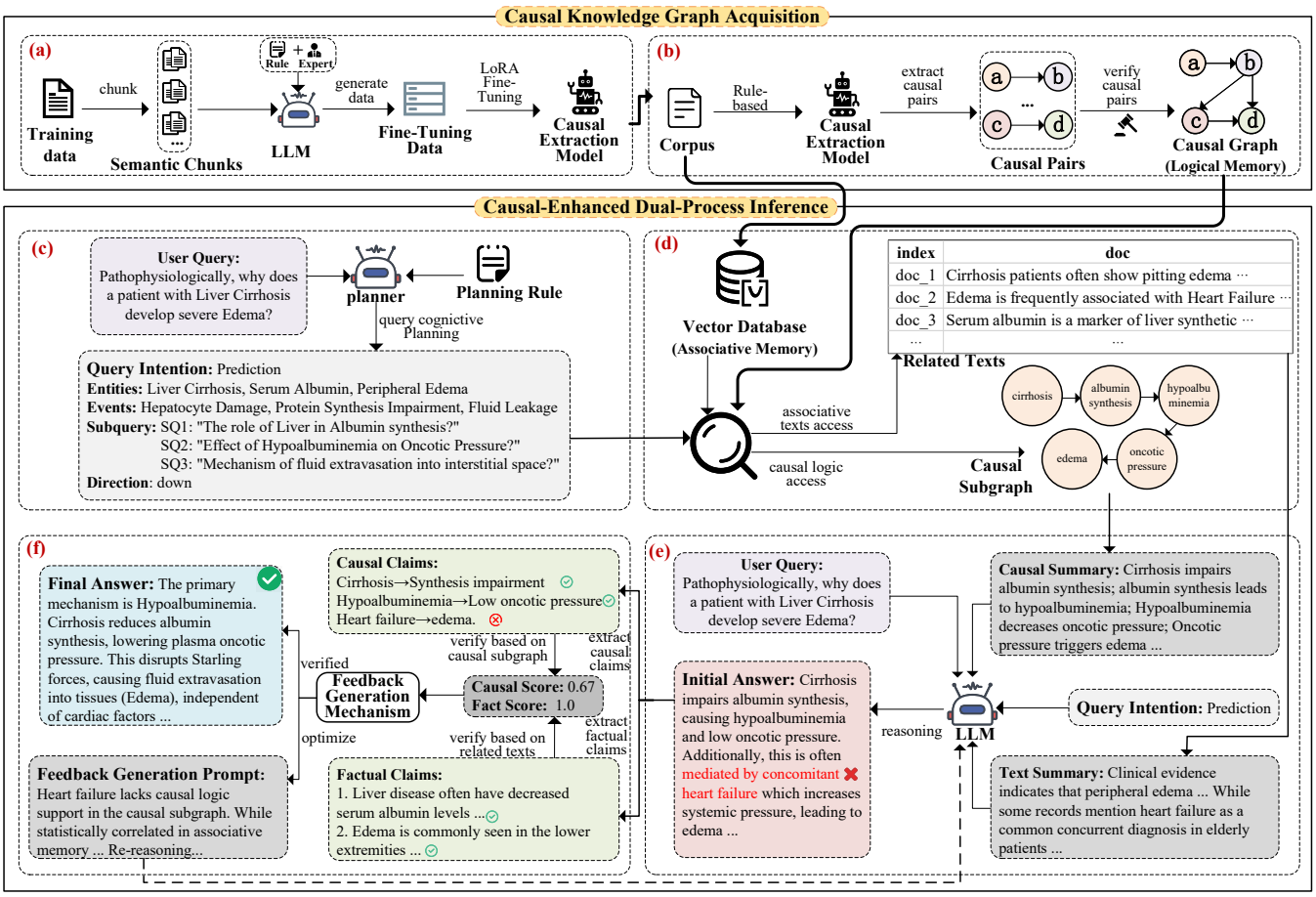

Figure 1: Overview of the Causal-Enhanced Dual-Process Inference framework. The acquisition of the Causal Knowledge Graph, features expert-supervised LoRA fine-tuning and a dual-verification mechanism for extracting reliable causal pairs from unstructured documents. The inference process, comprising (c) Query Cognitive Planner (intent analysis), (d) Dual-Path Memory Access (synergizing associative vector memory and structured causal logical memory), (e) Answer Generation, and a "System 2" inspired (f) Metacognitive Validation loop that optimizes responses using causal and factual feedback.

tion. This process ensures the graph captures complex logical dependencies while minimizing noise.

Semantic Chunking and Data Preparation The raw corpus is first segmented into Semantic Chunks to preserve the contextual integrity of causal events. To build a robust instruction-tuning dataset, we employ an expert-supervised data generation strategy. We prompt a general-purpose Large Language Model (LLM) using Few-Shot Prompts and predefined Causal Extraction Rules to generate initial causal candidates. Domain experts then rigorously verify these candidates, filtering out hallucinations and logically unsound pairs. This results in a high-quality dataset $\mathcal{D}_{train} = \{(x_i, y_i)\}$, where x_i represents the context and instructions, and y_i represents the ground-truth causal pairs verified by experts.

LoRA Fine-Tuning for Causal Extraction We employ Low-Rank Adaptation (LoRA) to fine-tune the base model by freezing pre-trained weights and injecting trainable rank decomposition matrices into the Transformer layers. This yields a specialized LoRA fine-tuned model capable of adhering to complex causal logic with high parameter efficiency.

Causal Graph Acquisition with Dual-Verification The fine-tuned model processes documents to extract causal pairs. Crucially, to ensure consistency, we inject the same Few-Shot Prompts and Extraction Rules used during training as guidance. The extracted pairs undergo a secondary automated verification step. A separate LLM serves as an adjudicator to validate the rationality of each pair. Only pairs that pass this verification are admitted into the final graph. This ensures the graph represents a reliable "long-term logical memory" for the system. The resulting Causal Knowledge Graph is formalized as a Directed Acyclic Graph (DAG), which prevents circular logic, ensuring valid downstream reasoning.

Causal-Enhanced Dual-Process Inference

The inference phase models the cognitive process of complex problem-solving, involving executive planning, dual-

Algorithm 1: Causal-Enhanced Dual-Process Inference

Input: User Query Q, Causal Graph $\mathcal{G}$, Vector Database DB, Thresholds τ_c, τ_f
Output: Final Answer A_{final}

$I \leftarrow \text{ClassifyIntent}(Q)$;
$d \leftarrow \text{MapDirection}(Q)$;
$Q_{sub}, E, V_{anchor} \leftarrow \text{Decompose}(Q)$;
$\mathcal{T}_{cand} \leftarrow \text{ANN}(\text{Embed}(Q_{sub} \cup E), DB)$;
$\mathcal{T}_{evidence} \leftarrow \text{ReRank}(Q_{sub}, \mathcal{T}_{cand})$;
$V_{ret} \leftarrow \bigcup_{v \in V_{anchor}} \{u \mid \text{dist}(v, u \mid d) \leq k\}$;
$\mathcal{G}_{sub} \leftarrow \text{Subgraph}(V_{ret}, \mathcal{G})$;
$S_{causal} \leftarrow \text{Summarize}(\mathcal{G}_{sub})$;
$S_{text} \leftarrow \text{Summarize}(\mathcal{T}_{evidence})$;
$iter \leftarrow 0$;
while $iter < Max_{iter}$ **do**
  $A_{final} \leftarrow \text{Generate}(Q, I, S_{causal}, S_{text})$;
  $C_c, C_f \leftarrow \text{ParseClaims}(A_{final})$;
  $Score_c \leftarrow \text{Verify}(C_c, \mathcal{G}_{sub})$;
  $Score_f \leftarrow \text{Verify}(C_f, \mathcal{T}_{evidence})$;
  if $Score_c \geq \tau_c \wedge Score_f \geq \tau_f$ **then**
    return A_{final};
  $P_{fb} \leftarrow \text{GenerateFeedback}(C_c, C_f)$;
  $A_{final} \leftarrow \text{Optimize}(A_{final}, P_{fb})$;
  $iter \leftarrow iter + 1$;
return A_{final};

path memory retrieval, and metacognitive self-correction. An overview can be found in Algorithm 1.

Query Cognitive Planner To address the ambiguity often present in natural language queries, the Query Cognitive Planner functions as a executive control module. Upon receiving a user query, an LLM first classifies its underlying cognitive intent into one of three categories: Attribution, which seeks causes or means (necessitating upstream reasoning from effect to cause); Prediction, which forecasts consequences or future trends (requiring downstream reasoning from cause to effect); or Definition, which requests concepts or holistic context (involving bidirectional exploration). Crucially, this intent classification dictates the reasoning direction for the "System 2" causal graph traversal (up, down, or both), filtering out irrelevant logical paths. Simultaneously, the module decomposes the complex query into independently retrievable sub-queries and extracts key entities and events, which serve as anchors for the subsequent memory access.

Dual-Path Memory Access We employ a dual-path strategy to synergize System 1 (Associative) and System 2 (Logical) retrieval processes.

Associative Retrieval (System 1). To capture semantic nuances, we access the "Associative Memory" (Vector Database). We first embed both the sub-queries (Q_{sub}) and extracted entities (E) into a shared vector space, and retrieve a broad set of candidate texts $\mathcal{T}_{cand}$ using Approximate Nearest Neighbor (ANN) search based on cosine similarity. To filter noise, a Re-ranking Model scores these candidates, selecting the top-k evidences $\mathcal{T}_{evidence}$:

$$\mathcal{T}_{cand} = \text{ANN}(\text{Embed}(Q_{sub} \cup E), \text{DB}, k_1) \tag{1}$$

$$\mathcal{T}_{evidence} = \text{TopK}(\text{ReRank}(Q_{sub}, \mathcal{T}_{cand}), k_2) \tag{2}$$

where k_1 is the initial retrieval window and k_2 is the refined window size.

Logical Retrieval (System 2). Simultaneously, we perform structure-aware retrieval on the "Logical Memory" (Causal Knowledge Graph) to capture causal logical dependencies. The Causal Knowledge Graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the universe of event nodes and $\mathcal{E}$ represents the set of causal edges.

Based on the events extracted during the Query Cognitive Planner, we identify a set of Anchor Nodes $V_{anchor} \subset \mathcal{V}$. The retrieval process is constrained by the determined direction $d \in \{\uparrow, \downarrow, \updownarrow\}$ and a maximum traversal depth k.

We define the set of retrieved nodes V_{ret} as the union of logical neighborhoods for all anchor nodes:

$$V_{ret} = \bigcup_{v \in V_{anchor}} \{u \in \mathcal{V} \mid \text{dist}(v, u \mid d) \leq k\} \tag{3}$$

where $\text{dist}(v, u \mid d)$ is the shortest path distance between node v and node u strictly following the direction d .

This process yields a causal subgraph $\mathcal{G}_{sub}$ that provides the necessary logical scaffolding for the reasoning task, operating independently of semantic similarity. $\mathcal{G}_{sub}$ is defined as the subgraph induced by these nodes:

$$\mathcal{G}_{sub} = \{(u, r, w) \in \mathcal{E} \mid u \in V_{ret} \wedge w \in V_{ret}\} \tag{4}$$

Answer Generation The retrieved contexts are synthesized into distinct summaries to mitigate information overload. The subgraph $\mathcal{G}_{sub}$ is abstracted into a structured Causal Summary (S_{causal}), while the text evidence $\mathcal{T}_{evidence}$ is condensed into a Text Summary (S_{text}). The LLM generates the initial answer A_{init} by conditioning on the query intent I and these summaries:

$$A_{init} = \arg\max_{y} P_\theta(y \mid Q, I, S_{causal}, S_{text}) \tag{5}$$

Metacognition Validation via Feedback Loop To ensure the response is logically sound and factually grounded, we introduce a "System 2" monitoring mechanism.

The initial answer is parsed into a set of Causal Claims (C_c) and Factual Claims (C_f). An LLM-as-a-Judge evaluates each claim against its respective source, calculating the causal score($Score_c$) and the fact score($Score_f$):

$$Score_c = \frac{1}{|C_c|} \sum_{c \in C_c} I(c \text{ entails } \mathcal{G}_{sub}) \tag{6}$$

$$Score_f = \frac{1}{|C_f|} \sum_{f \in C_f} I(f \text{ entails } \mathcal{T}_{evidence}) \tag{7}$$

where $I(\cdot)$ is the verification function returning 1 if supported, else 0.

Finally, the feedback generation mechanism aggregates these scores. We define validity thresholds τ_c and τ_f. If the answer fails to meet these standards, a feedback prompt P_{fb} is generated, explicitly listing the unsupported claims to guide optimization. The final answer A_{final} is determined by:

$$A_{final} = \begin{cases} A_{init} & \text{if } Score_c \geq \tau_c \wedge Score_f \geq \tau_f \\ \text{Optimize}(A_{init}, P_{fb}) & \text{otherwise} \end{cases} \tag{8}$$

To prevent infinite reasoning loops, we enforce a maximum iteration limit (Max_{iter}), defaulting to the locally optimal answer. This feedback loop as a cognitive intervention that effectively mitigates hallucinations and ensures the final output aligns with both causal logic and textual facts.

Experiment

Experimental Setting

Datasets. We evaluated our cognitive framework on three datasets representing distinct reasoning challenges: **Ship Intelligence Dataset (ShipInt)**, a self-constructed maritime corpus testing implicit causal reasoning in a sparse-data domain, requiring strong System 2 logic to bridge gaps; **MedQA** (D. Jin et al., 2021), requiring rigorous biomedical deduction, where associative errors can be critical; and **HotpotQA** (Yang et al., 2018), serving as a testbed for the interaction between associative breadth and logical depth. Particularly, The LoRA fine-tuning utilized only the training splits of the respective corpora.

Models. For Causal Graph Acquisition, we used **Qwen3-8B** with LoRA fine-tuning for efficient rule internalization. To demonstrate model-agnostic generalizability, we employ **Qwen3-14B** as a mid-sized reasoner and **DeepSeek-v3.2** (Liu et al., 2025) as the state-of-the-art upper bound.

Baselines. We benchmarked our proposed framework against four paradigms, treating them as proxies for different memory models. **Associative Memory Only (Naive-RAG)**, simulates purely associative memory using dense vector retrieval without structure; **Undirected Relational Memory (GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2024))**, captures global structural context but lacks the directed causality required for deductive reasoning; and **Associative-Structural Hybrid (HippoRAG2 (Gutiérrez et al., 2025))**, combines PageRank with knowledge graphs, representing a sophisticated associative model but lacking explicit causal reasoning mechanisms.

Metrics. We evaluate Causal Graph Acquisition quality using **BLEU-4** and **ROUGE-1/2/L**. For Reasoning Performance, we report **Context Recall** (retrieval effectiveness) and **Answer Correctness** (end-to-end reasoning fidelity assessed by Ragas framework(Es, James, Anke, & Schockaert, 2024)).

Table 1: Impact of LoRA Fine-Tuning on Causal Extraction Quality.

Configuration	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
No Fine-Tuning	41.13	67.32	41.89	45.84
Ours	**69.16**	**83.63**	**77.44**	**77.19**
Improvement	*+28.03*	*+16.31*	*+35.55*	*+31.35*

Table 2: Answer Correctness scores across three datasets using Qwen3-14B and DeepSeek-v3.2 backbones.

Dataset	ShipInt	MedQA	HotpotQA
Qwen3-14B			
Naive-RAG	79.96	79.26	40.97
GraphRAG	83.37	81.31	52.23
LightRAG	83.85	81.50	41.15
HippoRAG2	88.13	88.96	59.36
Ours	**89.97**	**92.58**	**63.72**
DeepSeek-v3.2			
Naive-RAG	82.53	84.53	49.26
GraphRAG	81.20	83.17	52.49
LightRAG	84.47	84.92	41.30
HippoRAG2	88.14	90.10	62.08
Ours	**91.03**	**95.56**	**64.73**

Results of Causal Graph Acquisition

To validate the effectiveness of our Causal Graph Acquisition strategy, we compared the performance of the Qwen3-8B backbone in a zero-shot setting (No Fine-Tuning) versus our fine-tuned setting. Table 1 presents the comparative results on the causal extraction task. The base model performs sub-optimally, particularly in ROUGE-2 (41.89) and BLEU-4 (41.13), indicating that generalist LLMs struggle to capture the precise syntactic and logical dependencies required for domain-specific causal extraction. In contrast, LoRA Fine-Tuning triggers a dramatic performance leap, with BLEU-4 surging to 69.16 and ROUGE-2 nearly doubling to 77.44.

Comparative Analysis of Reasoning Architectures

We evaluated the Dual-Process Inference performance across two dimensions: reasoning fidelity (Answer Correctness) and memory search effectiveness (Context Recall).

Results in Table 2 reveal a consistent "Logical Advantage". Our Dual-Process framework consistently outperforms purely associative baselines. With DeepSeek-v3.2, we achieve state-of-the-art results on ShipInt (91.03) and MedQA (95.56). Notably, our method using the smaller Qwen3-14B achieves 89.97 on ShipInt, significantly outperforming the larger DeepSeek-based Naive-RAG (82.53). As

Table 3: Context Recall scores using the DeepSeek-v3.2 backbone.

Dataset	ShipInt	MedQA	HotpotQA
Naive-RAG	38.85	19.31	56.72
GraphRAG	57.01	70.37	34.67
LightRAG	87.17	73.86	55.45
HippoRAG2	<u>88.39</u>	<u>84.65</u>	**96.30**
Ours	**92.33**	**87.13**	<u>81.76</u>

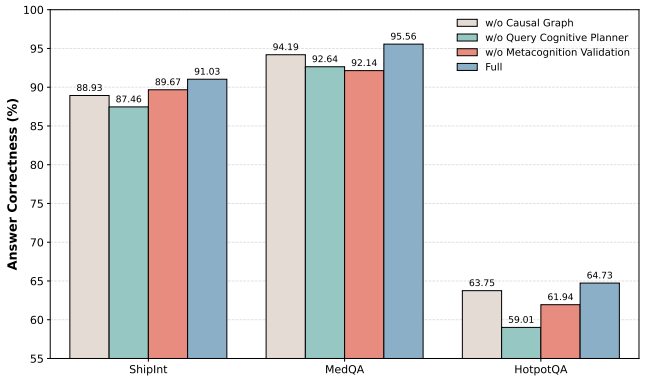

Figure 2: Ablation study quantifying the contribution of individual modules.

shown in Table 3, we achieve the highest recall on ShipInt (92.33) and MedQA (87.13), but falls behind HippoRAG2 on HotpotQA (81.76 vs. 96.30).

Ablation Study

To verify the contribution of specific components, we conducted ablation tests using the DeepSeek-v3.2 backbone, with results in Figure 2. Removing Query Cognitive Planner module significantly degraded HotpotQA (59.01) and ShipInt (87.46) performance, confirming the necessity of explicit cognitive planning. The Metacognition Validation module proved vital for high-precision domains like MedQA (dropping to 92.14), validating the necessity of a "System 2" monitor to mitigate subtle hallucinations. Finally, removing the Causal Graph caused regression across all datasets, validating its role as a critical "Logical Scaffold" for deductive reasoning that prevents reliance on shallow associative heuristics.

Parameter Analysis

To optimize the cognitive economy of our framework, we analyzed the sensitivity of the feedback generation mechanism by varying the Causal (τ_c) and Factual (τ_f) thresholds within the Metacognitive Validation module. As presented in Table 4, the results illustrate a clear trade-off between reasoning fidelity and computational cost. Under loose constraints ($\tau \approx 0.6$), the system minimizes latency (41.7s) but sacrifices accuracy (89.76%) as it fails to intercept subtle hallucinations. The Default Setting ($\tau_c = 0.75$,$\tau_f = 0.8$) achieves the optimal equilibrium, peaking at 91.03% correctness with manageable latency (47.92s). Notably, settings of High Strictness trigger a state of diminishing returns, while latency spikes to 58.22s due to frequent iterative loops, accuracy plateaus, indicating that hyper-correction introduces computational overhead without yielding further logical gains.

Table 4: Parameter Analysis of Verification Thresholds. We evaluate the trade-off between Answer Correctness and Inference Latency on the ShipInt dataset.

Setting Description	τ_c	τ_f	Answer Correctness	Latency (s)
Loose Constraints	0.6	0.7	89.76	41.71
Factual Focus	0.7	0.85	90.21	50.49
Default Setting	0.75	0.8	91.03	**47.92**
Strict Parity	0.8	0.85	**91.05**	52.78
High Strictness	0.9	0.9	90.98	58.223

Discussion

Our study validates that integrating causal graphs as "logical scaffolding" and employing Metacognitive Monitoring for "System 2" self-correction outperforms purely associative baselines in deductive tasks (e.g., MedQA). However, a cognitive trade-off emerged: directional logical retrieval restricts the associative breadth needed for broad multi-hop exploration (lower recall on HotpotQA). The parameter analysis highlights the cost of "slow thinking". While stricter thresholds improve accuracy, the "Hyper-Strict" setting yields diminishing returns with latency, confirming that an optimal architecture must balance verification rigor with computational efficiency. Our framework alleviates the pressure of large-scale human annotation. By initially fine-tuning a model via LoRA with minimal expert data in a general domain, this model can subsequently serve as a "Teacher" to automate high-fidelity causal graph extraction in new target domains. Future work will develop an Adaptive Cognitive Planner to dynamically balance this logical precision with associative breadth based on query complexity.

Conclusion

In this paper, we introduced a cognitive framework that aligns LLM reasoning with human-like Dual-Process cognition. By synergizing associative memory with a high-fidelity Causal Knowledge Graph, our approach effectively mitigates the logical shallowness of standard retrieval-based methods. Empirical results confirm that our Metacognitive Validation module acts as a robust "System 2" monitor, significantly reducing hallucinations in complex deductive tasks. While our directional logical retrieval ensures high precision, it presents a trade-off with exploratory breadth. Future work will address this by developing an Adaptive Cognitive Planner to dynamically balance logical strictness with associative exploration, moving closer to flexible, human-like reasoning.

Acknowledgments

Our work is supported by Dongguan Key Research & Development Program, China (No. 20241200300122), Guangdong Basic and Applied Basic Research Foundation (2025A1515011306), and Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

References

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations.*

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for ai. *Communications of the ACM*, *64*(7), 58–65.

Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., ... Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, *37*, 96640–96670.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning.*

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations* (pp. 150–158).

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223–241.

Fu, Z., Bao, G., Zhang, H., Hu, C., & Zhang, Y. (2025). Correlation or causation: Analyzing the causal structures of llm and lrm reasoning process. *arXiv preprint arXiv:2509.17380*.

Gao, J., Ding, X., Qin, B., & Liu, T. (2023). Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., ... others (2023). Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 16477–16508).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1).

Garcez, A. d., & Lamb, L. C. (2023). Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, *56*(11), 12387–12406.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2023). Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., & Su, Y. (2025). From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929–3938).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... others (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, *11*(14), 6421.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., ... others (2023). Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, *36*, 31038–31065.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 6769–6781).

Kiciman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459–9474.

Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., ... others (2025). Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Long, S., Schuster, T., & Piché, A. (2023). Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the association for computational linguistics: Emnlp 2023* (pp. 9248–9274).

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal rein-

forcement learning. *Advances in Neural Information Processing Systems*, *36*, 8634–8652.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., … others (2024). Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the association for computational linguistics: Acl 2024* (pp. 13697–13720).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Xiong, W., Li, X. L., Iyer, S., Du, J., Lewis, P., Wang, W. Y., … others (2020). Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on empirical methods in natural language processing (EMNLP)*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, *36*, 11809–11822.

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., … others (2025). siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, *51*(4), 1373–1418.