CC-RAG: Structured Multi-Hop Reasoning via Theme-Based Causal Graphs

Anonymous ACL submission

Abstract

Understanding cause and effect relationships remains a formidable challenge for Large Language Models (LLMs), particularly in specialized domains where reasoning requires more than surface-level correlations. Retrieval-Augmented Generation (RAG) improves factual accuracy, but standard RAG pipelines treat evidence as flat context, lacking the structure required to model true causal dependencies.

We introduce **Causal-Chain RAG (CC-RAG)**, a novel approach that integrates zero-shot triple extraction and theme-aware graph chaining into the RAG pipeline, enabling structured multi-hop inference. Given a domain specific corpus, CC-RAG constructs a Directed Acyclic Graph (DAG) of $\langle cause, relation, effect \rangle$ triples and uses forward/backward chaining to guide structured answer generation.

Experiments on two real-world domains: Bitcoin price fluctuations and Gaucher disease, show that CC-RAG outperforms standard RAG and zero-shot LLMs in chain similarity, information density, and lexical diversity. Both LLM-as-a-Judge and human evaluations consistently favor CC-RAG.

Our results demonstrate that explicitly modeling causal structure enables LLMs to generate more accurate and interpretable responses, especially in specialized domains where flat retrieval fails.

1 Introduction

Understanding and interpreting causal relationships embedded within specialized domains, such as healthcare and finance, is crucial to gain actionable insights. Although Large Language Models (LLMs) such as GPT (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2023) excel in tasks like question answering (Pichappan et al., 2023; Bahak et al., 2023), summarization (Ravaut et al., 2024), and localized information extraction (Yuan et al., 2024; Ding et al., 2024), they often fall short when tasked with chaining together information into coherent explanations (Zečević et al., 2023).

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Mansurova et al., 2024) improves LLM factuality by retrieving relevant context from external knowledge sources. However, standard RAG systems treat retrieved text as flat inputs, lacking the structure necessary to model true multi-hop dependencies. As a result, they often fail to explain nuanced insights behind connections, particularly in niche domains.

Classical causal inference methods focus on statistical associations and are not well equipped to extract narratives from unstructured text, especially when dependencies are implicit or span across documents (Weinberg et al., 2024; Cox and Wermuth, 2004; Hassani et al., 2017).

To bridge this gap, we introduce **CC-RAG**, a framework that explicitly integrates causal structuring into the RAG pipeline to support interpretable multi-hop reasoning. Given a domain-specific corpus, CC-RAG uses GPT-4o (Achiam et al., 2023) to extract zero-shot ⟨*cause, relation, effect*⟩ triples, which are then organized into a Directed Acyclic Graph (DAG). At inference time, the system performs both forward and backward chaining over this DAG to retrieve causal chains aligned to a given query, grounding the LLM's generation process with structured, traceable reasoning paths.

Our approach builds upon the ThemeKG framework (Ding et al., 2024), which constructs narrow-scope knowledge graphs focused on a specific domain. While ThemeKG improves entity salience and granularity compared to broad KGs like Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Speer et al., 2017), it does not model causality or support reasoning. CC-RAG fills this gap

[*] All code, datasets, prompts, and evaluation scripts will be released publicly to support full reproducibility.

by adding causal semantics and traversal capabilities, enabling fine-grained explanations grounded in domain-specific evidence.

We evaluate CC-RAG on two real-world datasets: **Bitcoin price fluctuations** (BP) (Wu et al., 2024), which requires modeling of volatile market dynamics, and **Gaucher disease** (GD) (Grabowski, 2008), which needs interpretation of biomedical causal pathways. Across both domains, CC-RAG consistently outperforms standard RAG and Zero-Shot LLMs in our automatic metrics, LLM-as-a-Judge studies, and human evaluations, demonstrating its ability to generate accurate and traceable multi-hop responses.

2 Related Work

Causal reasoning remains a central challenge in artificial intelligence (Yamin et al., 2024; Joshi et al., 2024). While classical statistical methods such as *Granger Causality* (Granger, 1969) and *Propensity Score Matching* (Dehejia and Wahba, 2002) have been effective in structured data settings, they are not designed to extract or reason over causal relationships embedded in unstructured text.

Causal Relation Extraction in NLP. Early causal relation extraction efforts relied on rule-based systems to detect explicit causal markers (e.g., "because", "due to") (Girju, 2003). More recent transformer-based methods such as CausalBERT (Li et al., 2021) and attention-enhanced relation extractors (Liu et al., 2019a; Li et al., 2019) improve performance by learning contextual representations. However, these models typically treat causality as pairwise classification and fail to capture long-range dependencies or build structured causal pathways, limiting their interpretability and reasoning capabilities.

ML-Based Causal Inference. Neural approaches to causal inference, such as DECI (Geffner et al., 2022), integrate causal discovery and effect estimation in end-to-end pipelines using GNNs and counterfactual reasoning (Pearl, 2009; Schölkopf et al., 2021). While effective on structured inputs, these models are computationally intensive and assume access to clean data, making them poorly suited for reasoning over raw text or document corpora.

Causal Graph Construction and Reasoning. Several recent methods build causal graphs from text. ERGO (Chen et al., 2022) classifies event pairs as causal or not to infer sparse links. QA systems like Think-on-Graph (Sun et al., 2023) and TIARA (Shu et al., 2022) retrieve subgraphs over large-scale KGs for factual questions. These approaches, however, do not model directional causality or construct explanatory chains, and therefore cannot support interpretable reasoning over cause-effect sequences.

Knowledge Graph Construction with LLMs. Early KGs such as ConceptNet (Speer and Lowry-Duda, 2017) and WordNet (Miller, 1995) encode broad semantic or taxonomic relationships but lack fine-grained semantics. More recent works (Chen et al., 2023; Fang et al., 2024; Li et al., 2024; Trajanoska et al., 2023; Wei et al., 2024; Zhu et al., 2024; Kau et al., 2024; Jiang et al., 2024, 2025) have explored LLMs for KG construction and completion. ThemeKG (Ding et al., 2024), in particular, narrows the focus of a KG to a single domain (e.g., finance, medicine), improving entity salience and relevance. However, ThemeKG does not extract causal relationships or support graph-based inference, limiting its use for tasks requiring explanation and reasoning.

Retrieval-Augmented Generation (RAG). Standard RAG frameworks improve factuality by conditioning LLM outputs on retrieved documents (Lewis et al., 2020). However, most treat retrieved evidence as unstructured context and lack any mechanism to represent or reason over causal structure. While multi-hop retrieval models (Asai et al., 2019) enable long-form question answering, they are not designed to trace cause-effect dependencies or compose interpretable reasoning chains.

Our Contributions. While prior work has explored causal extraction, KG construction, and retrieval-based generation in isolation, none integrate all three to support structured causal reasoning over raw text corpora. We introduce **CC-RAG**, a unified framework that combines zero-shot causal triple extraction, domain-specific DAG construction, and multi-hop retrieval-guided generation.

Zero-Shot Extraction of Causal Triples. CC-RAG uses GPT-4o to extract both explicit and implicit $\langle$*cause, relation, effect*$\rangle$ triples from unstructured text, without requiring labeled training data.

Theme-Aligned Causal KGs. We extend ThemeKG by organizing causal triples into Directed Acyclic Graphs that capture fine-grained dependencies, enabling interpretable and domain-aligned graph traversal.

Bidirectional Multi-Hop Reasoning. Unlike previous systems that rely on single-hop evidence or unstructured retrieval, CC-RAG supports both forward and backward chaining over its causal graph, tracing from causes to effects or vice versa.

Together, these innovations enable CC-RAG to serve as a scalable, interpretable reasoning layer over unstructured corpora, outperforming traditional methods across our automated metrics, LLM-as-a-Judge evaluations, and human evaluations.

3 CC-RAG

The CC-RAG framework, as described in Figure 1, is introduced below.

3.1 Entity Extraction and DAG Construction

CC-RAG uses GPT-4o to extract ⟨*cause, relation, effect*⟩ triples from our text corpora. We use a zero-shot prompting strategy described below to identify both explicit (word-based) and implicit (an inferred connection that can be understood contextually) causal links, ensuring a more comprehensive representation of causal relationships. When applicable, the extracted entities and relations are aligned with domain-specific taxonomies (e.g., **MeSH** (National Library of Medicine, 1960) for biomedical datasets) to enhance consistency. Given a document, the model is prompted to extract causal triples, where:
Cause refers to an entity, event, or action that triggers an outcome, even if the causal connection is not explicitly stated.
Relation is a causal verb or phrase (e.g., *caused, led to, resulted in, triggered, influenced*), or an inferred connection that is understood contextually.
Effect represents the resulting entity, event, or action, regardless of whether the causal relationship is directly stated in the text.

These extracted triples form the basis of a **Directed Acyclic Graph (DAG)** that captures both direct and indirect causal dependencies, enabling structured reasoning over the text.

3.2 User Query Processing

Given a user query, we extract key entities and concepts using *KeyBERT* (Grootendorst, 2020), a transformer-based keyword extraction to identify the most relevant words and phrases from the query while filtering out stop words and irrelevant terms. These extracted keywords and phrases are matched against nodes in the KG using *Sentence-BERT* (Reimers, 2019) embeddings, allowing semantic similarity-based retrieval of relevant entities.

The user query's causal direction is classified as *forward* (identifying effects) or *backward* (identifying causes) using an LLM prompt. This classification ensures that the retrieval process aligns with the user's intent. The extracted key entities are used to retrieve the relevant causal pathways from the KG via forward or backward chaining depending on the identified causal direction.

3.3 Graph-Based Multi-Hop Reasoning

The LLM enhanced querying strategy on the DAG supports:
Forward Chaining: Starting from an initial cause, traverse downstream effects to construct a causal sequence.
Backward Chaining: Given an observed effect, trace upstream paths to uncover potential causes.

This bidirectional reasoning allows CC-RAG to generate answers that not only respond to the query but also provide structured, explainable causal pathways.

3.4 LLM-Guided Chain Ranking

Given a user query, we identify semantically relevant nodes in the theme-KG and retrieve multiple forward or backward causal chains from these entry points. To rank the candidate chains retrieved, the LLM is prompted with all chains and asked to select the one that best answers the query. This enables adaptive chain selection based on semantic alignment, logical coverage, and answer completeness. Prior studies have shown that LLMs excel in ranking tasks by leveraging vast contextual embeddings to assess *semantic similarity* and *coherence* (Mann et al., 2020; Raffel et al., 2020; Zhuang et al., 2023; Hou et al., 2024; Qin et al., 2023). Unlike heuristic-based methods, our framework does not use hard-coded thresholds. Instead, the LLM dynamically selects the most contextually appropriate chain, enabling flexible adaptation across domains. To promote diversity and avoid overfitting to surface-level paths, chains are generated from multiple matched nodes, ensuring coverage of distinct subgraphs and allowing the model to surface less obvious but semantically relevant causal pathways.

3.5 Output Generation with Justification

The final step synthesizes the reasoning process into a natural language response using LLM-powered summarization and explanation. Given

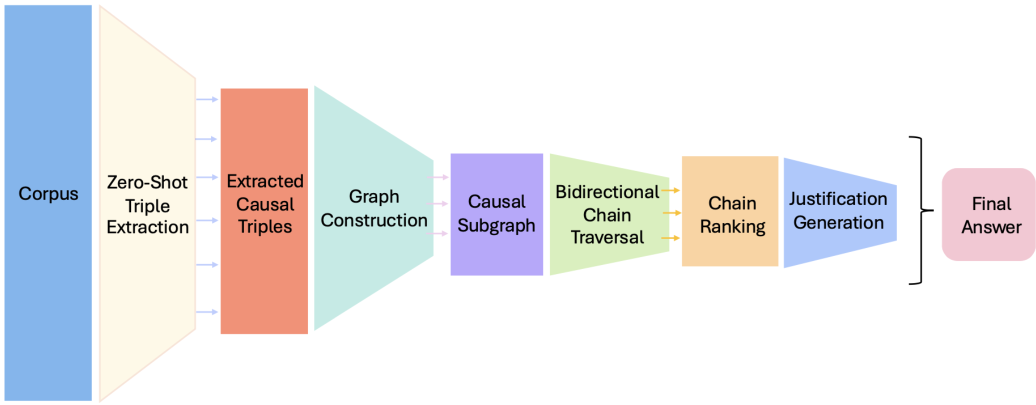

Figure 1: The CC-RAG framework. From a domain-specific corpus, GPT-4o extracts zero-shot causal triples, which are structured into a DAG. Given a query, the system identifies semantic matches, performs forward or backward traversal to extract causal chains, and generates a justification-based answer using an LLM. This architecture enables interpretable multi-hop reasoning.

the selected causal pathway, the LLM generates a structured response that includes a direct answer to the user's query, grounded in the retrieved causal chain and a justification explaining the reasoning process, explicitly tracing the multi-hop causal pathway identified in the knowledge graph. By leveraging an LLM for both structured response generation and justification, our approach improves trustworthiness and interpretability in AI-driven causal analysis.

A qualitative case study illustrating this process and comparing CC-RAG's zero-shot extraction capabilities vs. a human annotator is shown in Figure 2. In this figure, CC-RAG extracts and recovers a multi-hop causal pathway that cannot be reached by the human annotated triples.

4 Experiments

4.1 Dataset Collection

We constructed two datasets, **Bitcoin Price Fluctuations (BP)** in the finance domain and **Gaucher Disease (GD)** in the medical domain. These datasets are sourced from a combination of news articles from reputable sources (e.g. Reuters, Financial Times, etc.) and scientific papers from PubMed and the National Institute of Health (NIH). The documents were selected based on their relevance, reliability, granularity, and diversity.
We collected and processed 20 high-quality documents in each dataset and extracted entities and causal triples from the text using GPT-4o as described in Section 3.1. The final dataset statistics are shown in Table 1.

Dataset	Documents	Entities	Triples
BP	20	317	172
GD	20	573	419

Table 1: Dataset statistics for each domain used in our experiments.

4.2 Baselines

We compare CC-RAG against two baselines.
GPT-4o with RAG: A retrieval-augmented system that grounds GPT-4o using locally retrieved document chunks. We implemented this by leveraging structured chunking (Lewis et al., 2020) for efficient segmentation and a retrieval mechanism that dynamically selects the most relevant context for each query, ensuring accurate and context-aware responses.
Zero-Shot GPT-4o: A baseline that employs GPT-4o out of the box without any augmentation.

4.3 Evaluation

Automatic Evaluation. The performance of the three systems was evaluated using three key metrics that capture different aspects of response quality:

Chain Similarity (BERTScore): To evaluate how closely the generated causal chain aligns with a gold-standard reference, we compute the F1 variant of BERTScore (Zhang et al., 2019), which measures token-level semantic similarity using contextual embeddings from a pre-trained language model (RoBERTa-large) (Liu et al., 2019b). Reference chains were manually curated from literature

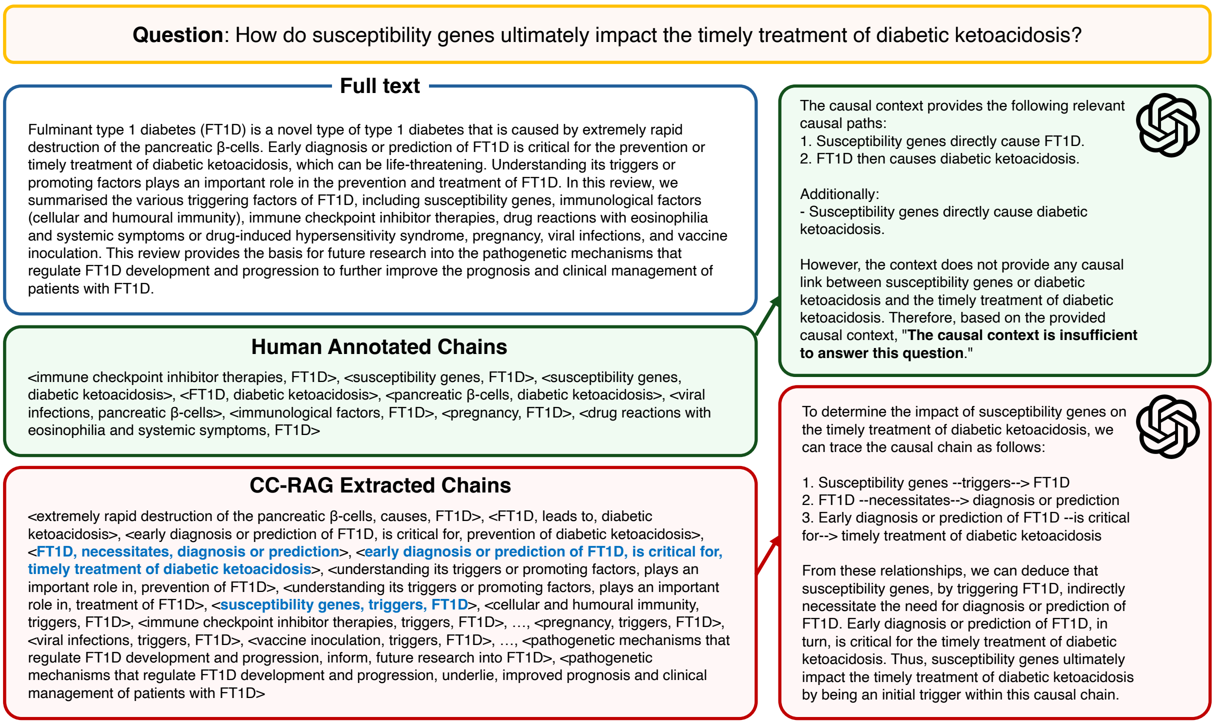

Figure 2: Qualitative case study comparing CC-RAG with expert-annotated causal triples (Antonucci et al., 2023) for answering a biomedical question. CC-RAG successfully reconstructs a multi-hop causal pathway, which the human-annotated KG fails to recover. This illustrates CC-RAG's ability to surface-latent causal connections via chain-based reasoning.

to represent the most widely accepted explanations for each question. A higher BERTScore indicates stronger semantic alignment between the generated and reference chains, suggesting that the model has effectively captured the key causal relationships.

Information Density (Conciseness): To measure how efficiently information is conveyed, we compute a normalized information density score as the ratio of content words to total words, scaled by the inverse log of response length. This metric serves as a proxy for conciseness, rewarding responses that are informative without excessive filler or redundancy.

Lexical Diversity (TTR): To assess linguistic variation in generated responses, we use the Type-Token Ratio (TTR) (Hess et al., 1984), defined as the number of unique words (types) divided by the total number of words (tokens). A higher TTR indicates greater lexical diversity, reflecting more varied word usage, while a lower TTR indicates repetition. In the context of text generation for QA, lexical diversity is necessary to produce concise yet accurate explanations without redundancy.

Overall Interpretation: High values across these metrics indicate that the system generates responses which are clear, semantically aligned with the intended reasoning, and succinct. Our goal is to ensure that the system not only provides correct and comprehensive information but does so in a manner that is easily traceable and free of irrelevant information.

LLM-as-a-Judge. To complement our automated metrics, we conducted a blinded LLM-as-a-Judge evaluation (Zheng et al., 2024) using a panel of four LLMs: GPT-4, GPT-4o, LLaMA 3.1-8B-Instruct, and Mistral-7B-Instruct. Each judge model was prompted with a fixed template and shown anonymized answers from all system (CC-RAG, LLM w/ RAG, and Zero-Shot LLM). Models were asked to select the best response based on accuracy, interpretability, and conciseness. We used majority voting to determine the preferred answer.

Human Evaluation. To further assess the quality of our generated responses, we conducted a human evaluation study for our case studies and model substitution experiment. Each study presented outputs from all systems for a given question, and asked reviewers to select the response that best answered the question, considering accuracy, interpretability, and conciseness. We aggregated the preferences of

Dataset	Metric	CC-RAG	RAG	Zero-Shot
		GPT-4o		
BP	BERTScore	**0.90**	0.73	0.00
	Conciseness	**16.44**	13.94	0.00
	TTR	**0.81**	0.73	0.00
GD	BERTScore	**0.88**	0.77	0.75
	Conciseness	**15.68**	13.20	11.66
	TTR	**0.74**	0.64	0.53
		LLaMA-3.1-8B-Instruct		
BP	BERTScore	**0.83**	0.71	0.00
	Conciseness	**12.04**	9.40	0.00
	TTR	**0.56**	0.40	0.00
GD	BERTScore	**0.85**	0.74	0.71
	Conciseness	**14.93**	10.28	10.40
	TTR	**0.69**	0.35	0.36

Table 2: Quantitative evaluation for GPT-4o and LLaMA-3.1-8B-Instruct.

three independent reviewers for each form to assess performance.

Infrastructure All of our experiments were conducted on Google Colab (Google, 2025), leveraging its GPU resources for our work. We estimate that we utilized approximately 75 GPU hours to execute all experiments.

5 Results

In this section, we assess performance by comparing key evaluation metrics across both datasets.

5.1 Automatic Quantitative Evaluation

Table 2 summarizes our evaluation metrics for two case studies. For both the BP and GD datasets, CC-RAG outperforms both GPT-4o with RAG and Zero-Shot GPT-4o across all metrics. These comprehensive improvements indicate that CC-RAG is more effective in providing detailed, precise, and contextually relevant responses.

5.2 LLM-as-a-Judge Evaluation Results

Table 3 summarizes the LLM-as-a-Judge results. CC-RAG consistently outperformed both GPT-4o with RAG and Zero-Shot GPT-4o across both BP and GD datasets, getting the majority vote in both cases. The "Split" votes, although limited, reflect some ambiguity from voting models, but CC-RAG remained the clear preferred system overall.

[*] All scores for Zero-Shot GPT-4o are 0 for the BP dataset because all queries referenced events that occurred outside of the LLM's training scope.

Evaluator	Dataset	CC-RAG	RAG	Zero-Shot	Split
		LLM-as-a-Judge			
LLM	BP	**14/18**	0/18	0/18	4/18
LLM	GD	**15/25**	3/25	2/25	5/25
		Human Evaluation			
Human	BP	**10/18**	5/18	0/18	3/18
Human	GD	**13/25**	6/25	2/25	4/25

Table 3: LLM-as-a-Judge and human evaluation results comparing CC-RAG, RAG, and Zero-Shot answers across both datasets. "Split" indicates disagreement among evaluators. CC-RAG was the most frequently preferred system in both domains.

5.3 Human Evaluation Results

Table 3 also summarizes our human evaluation results for the main experiments in both datasets. Across our datasets, CC-RAG was most frequently preferred by human evaluators, receiving the majority vote for each dataset. These results further validate CC-RAG's ability to generate human-interpretable and domain-relevant answers to complex queries.

5.4 Qualitative Case Study: Matching Expert-Annotated Chains

To better understand CC-RAG's multi-hop reasoning capabilities, we conducted a focused case study using a short biomedical paragraph from Antonucci et al. (2023), which includes a set of human-annotated causal triples. The paragraph describes the pathophysiology and progression of Fulminant Type 1 Diabetes (FT1D). From this paragraph, we constructed a complex causal query:

"How do susceptibility genes ultimately impact the timely treatment of diabetic ketoacidosis?"

We compare two methods for answering this question:

Human-Annotated KG + Zero-Shot Generation: We use the expert-annotated causal triples from Antonucci et al. to construct a knowledge graph. We then apply zero-shot prompting over this graph to generate an answer, simulating a best-case baseline with gold-standard edges but no structure-aware traversal.

CC-RAG: We apply our full CC-RAG pipeline to the same paragraph, performing zero-shot causal triple extraction, building a ThemeKG, performing backward chaining, and generating a justification-based answer.

As shown in Figure 2, the human-annotated KG fails to capture the full reasoning chain required to link susceptibility genes to treatment. In contrast, CC-RAG reconstructs the full multi-hop pathway:

```
susceptibility genes → FT1D → early
    diagnosis → timely treatment
```

This chain includes both direct and inferred causal links and enables the model to justify the final answer with a complete, interpretable path. The case study highlights that CC-RAG extracted a broader and more informative set of causal relationships from the text, and effectively structured them into a coherent and accurate reasoning chain.

5.5 Inference Efficiency

We compared average inference times between CC-RAG and traditional RAG across both case studies. Our results, as shown in Table 4 demonstrates that CC-RAG achieves faster response times across both domains. This improvement stems from CC-RAG's use of targeted KG traversal, which focuses the LLM's attention on concise and relevant paths. In contrast, traditional RAG retrieves broader document chunks that slow down inference.

Dataset	CC-RAG (s)	RAG (s)
BP	1.04	1.42
GD	1.10	2.41

Table 4: Average inference time (seconds) per query across datasets.

Overall, these results reinforce the capability of CC-RAG and suggest that integrating structured causal reasoning into RAG significantly improves answer interpretability and accuracy, particularly in specialized domains requiring multi-hop causal inference.

6 Generalization under Model Shift

We also conducted a model substitution experiment, in which we replaced the downstream analysis LLM (GPT-4o) with LLaMA-3.1-8B-Instruct. For this experiment, GPT-4o was still used for high quality entity extraction and causal graph construction, but all subsequent processing, including candidate chain analysis and selection, and summarization was performed using LLaMA. The objective here was to assess whether the structure provided by CC-RAG remains effective for reasoning, even with a less powerful model.

Evaluator	Dataset	CC-RAG	RAG	Zero-Shot	Split
LLM-as-a-Judge					
LLM	BP	**14/18**	1/18	2/18	1/18
LLM	GD	**13/25**	2/25	8/25	2/25
Human Evaluation					
Human	BP	**15/18**	3/18	0/18	0/18
Human	GD	**15/25**	4/25	1/25	5/25

Table 5: LLM-as-a-Judge and human evaluation results for the study using LLaMA. CC-RAG is compared against RAG and Zero-Shot baselines with the same underlying model. "Split" indicates evaluator disagreement. CC-RAG remained the preferred system across domains.

6.1 Quantitative Results

Table 5 summarizes the results of our model substitution experiment. Here, we compared three configurations: LLaMA CC-RAG, LLaMA with RAG, and Zero-Shot LLaMA, across both datasets.

Once again, CC-RAG outperformed both its RAG and Zero-Shot counterparts across all metrics in both datasets. This indicates that the responses are more relevant and focused, while remaining technically nuanced *.

6.2 LLM-as-a-Judge Evaluation

The LLM-as-a-Judge evaluations further support CC-RAG's superior performance even when a less powerful model like LLaMA is used, as shown in Table 5.

6.3 Human Evaluation

To underscore the impact of CC-RAG, we repeated our human evaluation under the model substitution setting. The same protocol was followed, with three annotators per question comparing CC-RAG, LLaMA with RAG, and Zero-Shot LLaMA. Results are shown in Table 5.

In this setting, CC-RAG saw an even stronger lead, receiving 83% of the votes in the BP study, and 60% of the votes in the GD study. This substantial gain relative to the RAG and Zero-Shot baselines highlights the importance of our structuring pipeline in producing accurate and interpretable causal answers.

The model substitution experiment highlights the value of integrating structured knowledge into an

LLM pipeline. It further demonstrates CC-RAG's ability to maintain high performance in resource constrained settings, making CC-RAG a promising solution for scalable causal reasoning applications.

7 Future Work

While CC-RAG demonstrates strong performance in structured causal reasoning, the following areas merit future exploration:

Multimodal Knowledge Integration Extending CC-RAG to incorporate images, graphs, and tabular data would enhance its applicability.

Adaptive Retrieval Mechanisms Developing dynamic retrieval mechanisms that update causal representations in real-time would provide relevant and up-to-date answers in rapidly evolving domains.

Scaling to Broader Domains We aim to extend CC-RAG to new domains including social sciences and legal reasoning.

By addressing these, CC-RAG can evolve into a more versatile and scalable framework, bridging the gap between knowledge-driven AI and real-world decision-making.

8 Limitations

We recognize that the CC-RAG framework has the following limitations:

LLM Dependency and Hallucinations CC-RAG relies on pre-trained LLMs for entity and relation extraction, which introduces potential hallucination and bias issues. LLMs may extract spurious causal links and reflect biases from training data. Future work could incorporate domain-specific fine-tuning to mitigate these effects.

Scalability of Graph Traversal As KGs grow, causal graph traversal can become computationally expensive, particularly for multi-hop reasoning. Optimizations such as graph pruning or embedding-based traversal would need to be implemented to enable scaling to large-scale applications.

Non-Determinism in LLM Inference LLM systems are inherently non-deterministic. As a result, executing CC-RAG multiple times even with the same query can yield varying results, potentially producing different answers on each run.

Summarization Drift Although prompting strategies are designed to guide LLM responses toward desired outcomes, the LLM may deviate from the intended chain, leading to hallucinations or inaccuracies.

Rapidly Evolving Space Recent advances in reasoning models such as OpenAI's o1 (OpenAI, 2024) and o3 (OpenAI, 2025), and DeepSeek's R1 (DeepSeek-AI et al., 2024) show promise in logical reasoning and causal inference. We plan to explore their effectiveness in knowledge graph integration and causal inference.

Despite these limitations, we believe that CC-RAG represents a significant step forward in structured causal reasoning. Addressing these limitations will further enhance CC-RAG's ability to provide robust, scalable, and interpretable causal insights.

9 Conclusion

In this paper, we introduced **CC-RAG**, a structured causal reasoning framework that integrates knowledge graphs with RAG to improve multi-hop causal inference. Unlike conventional RAG approaches, CC-RAG explicitly models causal pathways using Directed Acyclic Graphs (DAGs), enabling more coherent, interpretable, and structured causal reasoning.

Through evaluation on two datasets, **Bitcoin Price Fluctuations** and **Gaucher Disease**, we demonstrate that CC-RAG outperforms standard RAG and zero-shot LLM baselines both in terms of quantitative metrics and qualitative assessments using LLM-as-a-Judge and human evaluation. Our model substitution experiment further confirms that even with less powerful LLM, the structured causal reasoning in CC-RAG remains effective.

CC-RAG has significant implications for decision-making in high-stakes domains. In biomedicine, it can model disease pathways and drug interactions. In finance, it can trace macroeconomic and geopolitical influences. In policy and law, it enables causal tracing for regulatory analysis and compliance.

These findings highlight the importance of explicit causal modeling in LLM-driven reasoning. By integrating structured knowledge representations, CC-RAG moves beyond surface-level retrieval to produce explanations that are more interpretable, reliable, and well-suited for real-world decision making. As AI systems continue to evolve, the fusion of causal reasoning and structured knowledge retrieval will be essential to enable transparent, explainable, and trustworthy AI-driven insights.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alessandro Antonucci, Gregorio Piqué, and Marco Zaffalon. 2023. Zero-shot causal graph extrapolation from text via llms. *https://arxiv.org/abs/2312.14670*.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *arXiv preprint arXiv:1911.10470*.

Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, and Arefeh Kazemi. 2023. Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models. *arXiv preprint arXiv:2312.07592*.

Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason HD Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with llms. *arXiv preprint arXiv:2305.09858*.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. Ergo: Event relational graph transformer for document-level event causality identification. *arXiv preprint arXiv:2204.07434*.

David R Cox and Nanny Wermuth. 2004. Causality: a statistical view. *International Statistical Review*, 72(3):285–305.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

Linyi Ding, Sizhe Zhou, Jinfeng Xiao, and Jiawei Han. 2024. Automated construction of theme-specific knowledge graphs. *arXiv preprint arXiv:2404.19146*.

Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. 2024. Moltc: Towards molecular relational modeling in language models. *arXiv preprint arXiv:2402.03781*.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. 2022. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.

Google. 2025. Google colaboratory. `https://colab.research.google.com/`.

Gregory A Grabowski. 2008. Phenotype, diagnosis, and treatment of gaucher's disease. *The Lancet*, 372(9645):1263–1271.

Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Hossein Hassani, Xu Huang, and Mansi Ghodsi. 2017. Big data and causality. *Annals of Data Science*.

Carla W Hess, Kelley P Ritchie, and Richard G Landry. 1984. The type-token ratio and vocabulary performance. *Psychological Reports*, 55(1):51–57.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Pengcheng Jiang, Lang Cao, Cao Xiao, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. Kg-fit: Knowledge graph fine-tuning upon open-world knowledge. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jimeng Sun, and Jiawei Han. 2025. Ras: Retrieval-and-structuring for knowledge-intensive llm generation. *arXiv preprint arXiv: 2502.10996*.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. Llms are prone to fallacies in causal inference. *arXiv preprint arXiv:2406.12158*.

Amanda Kau, Xuzeng He, Aishwarya Nambissan, Aland Astudillo, Hui Yin, and Amir Aryani. 2024. Combining knowledge graphs and large language models. *arXiv preprint arXiv:2407.06564*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dawei Li, Zhen Tan, Tianlong Chen, and Huan Liu. 2024. Contextualization distillation from large language model for knowledge graph completion. *arXiv preprint arXiv:2402.01729*.

Pengfei Li, Kezhi Mao, Xuefeng Yang, and Qi Li. 2019. Improving relation extraction with knowledge-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 229–239, Hong Kong, China. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, Kuo Liao, Bing Qin, and Ting Liu. 2021. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision. *arXiv preprint arXiv:2107.09852*.

Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. 2019a. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 134–139, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

Aigerim Mansurova, Aiganym Mansurova, and Aliya Nugumanova. 2024. Qa-rag: Exploring llm reliance on external knowledge. *Big Data and Cognitive Computing*, 8(9):115.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

National Library of Medicine. 1960. Medical Subject Headings (MeSH). Accessed: February 15, 2025.

OpenAI. 2024. O1 system card.

OpenAI. 2025. Openai o3-mini system card.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Pit Pichappan, M Krishnamurthy, and P Vijayakumar. 2023. Analysis of chatgpt as a question-answering tool. *Journal of Digital Information Management*, 21(2):50–60.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting, 2023. *URL https://arxiv.org/abs/2306*, 17563.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. 2022. Tiara: Multi-grained retrieval for robust question answering over large knowledge bases. *arXiv preprint arXiv:2210.12925*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yanbin Wei, Qiushi Huang, James T Kwok, and Yu Zhang. 2024. Kicgpt: Large language model with knowledge in context for knowledge graph completion. *arXiv preprint arXiv:2402.02389*.

Abraham Itzhak Weinberg, Cristiano Premebida, and Diego Resende Faria. 2024. Causality from bottom to top: A survey. *arXiv preprint arXiv:2403.11219*.

Jingyang Wu, Xinyi Zhang, Fangyixuan Huang, Haochen Zhou, and Rohtiash Chandra. 2024. Review of deep learning models for crypto price prediction: implementation and evaluation. *arXiv preprint arXiv:2405.11431*.

Khurram Yamin, Shantanu Gupta, Gaurav R Ghosal, Zachary C Lipton, and Bryan Wilder. 2024. Failure modes of llms for causal reasoning on narratives. *arXiv preprint arXiv:2410.23884*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Jun Zhao, and Kang Liu. 2024. Improving zero-shot LLM reranker with risk minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17967–17983, Miami, Florida, USA. Association for Computational Linguistics.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*.

A Appendix

In this section, we provide specific examples of responses from CC-RAG, the LLM with RAG, and the Zero-Shot LLM from our experiments and the model substitution study.

A.1 Experiment Results and Analysis

Tables 6 and 7 show the responses from the CC-RAG and the baseline approaches to two sample queries from the BP dataset.
A qualitative analysis of BP Query 1 in Table 6 shows the following:

- **Strengths of CC-RAG:** *Plausible Inference and Contextual Reasoning:* CC-RAG generates a well-structured explanation that infers plausible causal links between Andrew Yang's advocacy and increased legitimacy of Bitcoin. Unlike GPT-4o with RAG, which fails to respond due to lack of directly retrieved mentions, CC-RAG integrates thematic understanding to bridge implicit gaps in the source context. Compared to Zero-Shot GPT-4o, CC-RAG delivers a domain-aware, relevant response rather than abstaining.

- **Weaknesses of CC-RAG:** *Potential Overgeneralization:* While CC-RAG fills contextual gaps effectively, it occasionally generalizes from weak evidence, which could lead to speculative claims if not properly validated against source content. Future iterations could integrate confidence calibration or provenance tracing to mitigate this.

A qualitative analysis of BP Query 2, found in Table 7 shows the following:

- **Strengths of CC-RAG:** *Clarity and Directness:* CC-RAG presents a straightforward explanation by directly linking global uncertainty to Bitcoin's price rise.

- **Weaknesses of CC-RAG:** *Lack of Nuanced Evidence Integration:* While the answer is topically accurate, it does not incorporate supporting details such as regulatory optimism or institutional inflows.

Tables 8 and 9 shows the responses from the CC-RAG and the baseline approaches to two sample queries from the GD dataset.
A qualitative analysis of GD Query 1, found in Table 8 shows the following:

- **Strengths of CC-RAG:** *Concise Causal Explanation:* CC-RAG provides a clear and direct answer that accurately captures the enzymatic deficiency and its consequence, glucocerebroside accumulation, while linking it to the observed symptoms of splenomegaly and hepatomegaly. Compared to GPT-4o with RAG, which includes detailed biological elaboration, CC-RAG prioritizes accessibility and interpretability. Unlike Zero-Shot GPT-4o, which offers comprehensive but sometimes diffuse explanations, CC-RAG delivers focused, cause-effect reasoning ideal for quick comprehension.

- **Weaknesses of CC-RAG:** *Lack of Cellular Detail:* CC-RAG doesn't include some biological intermediaries. GPT-4o with RAG and Zero-Shot GPT-4o both provide richer mechanistic insight, which can be more informative for expert audiences.

A qualitative analysis of GD Query 2, found in Table 9 shows the following:

- **Strengths of CC-RAG:** *Targeted Causal Reasoning:* CC-RAG delivers a precise explanation that links enzyme deficiency to macrophage-specific accumulation, accurately highlighting the role of phagocytic activity and impaired degradation. It maintains clarity while correctly identifying the cellular actors involved. Compared to GPT-4o with RAG, which emphasizes cell lineage bias, and Zero-Shot GPT-4o, which includes broader narrative background, CC-RAG focuses on the most relevant mechanistic cause.

- **Weaknesses of CC-RAG:** *Surface-Level Language:* The explanation lacks the technical specificity an expert might need. It avoids terminology like "glycosphingolipid metabolism," which could be useful for experts.

A.2 Model Substitution Study Results and Analysis

Table 10 and 11 show the responses from the CC-RAG and the baseline approaches for two sample queries from the BP dataset for the model substitution study.

Query	How did Andrew Yang's advocacy for blockchain technology influence Bitcoin's adoption?
CC-RAG	Andrew Yang's advocacy for blockchain technology brought increased public and political awareness to cryptocurrencies, such as Bitcoin. By discussing the potential benefits of blockchain, such as transparency and efficiency, Yang helped legitimize the technology in the eyes of the public and policymakers. This likely contributed to broader acceptance and interest in Bitcoin, although direct causation to adoption rates is harder to measure.
GPT-4o with RAG	The context provided does not mention Andrew Yang or his advocacy for blockchain technology. Therefore, there's no information on how Andrew Yang's advocacy might have influenced Bitcoin's adoption.
Zero-Shot GPT-4o	N/A

Table 6: Responses from different approaches for BP Query

Query	How did global economic uncertainty contribute to Bitcoin's price rise?
CC-RAG	Global economic uncertainty can lead to increased demand for Bitcoin as investors look for alternative assets that are perceived as safe havens. This heightened demand for Bitcoin drives up its price.
GPT-4o with RAG	Several contributing factors likely include: 1. **Speculative Investment**: The overall sentiment captured in the passage indicates that part of Bitcoin's price surge can be attributed to increased speculative investment, driven by optimism about a favorable regulatory environment and inflows of new investors, as detailed by the remarks from Chris Akhavan and Joel Kruger. 2. **Mainstream Adoption**: Another key influence is the increased mainstream adoption of Bitcoin, highlighted by the SEC's approval of new exchange-traded funds directly investing in cryptocurrency. This move allowed major global asset managers like BlackRock Inc. to participate, signaling institutional confidence and involvement in the market. 3. **Market Momentum**: Bitcoin's historic performance and recognition, compared to assets like gold and influenced by significant market movements and high-profile endorsements, have contributed to its rising status as a major asset class. Overall, while the context does not provide a direct explanation of how global economic uncertainty played a role, the speculative nature of investments and the moves towards mainstream adoption in a possibly volatile economic environment could be seen as indirect contributors to Bitcoin's price rise.
Zero-Shot GPT-4o	N/A

Table 7: Responses from different approaches for BP Query 2

A qualitative analysis of Model Substitution BP Query 1, found in Table 10 shows the following:

- **Strengths of CC-RAG:** *Resilient Inference Under Model Substitution:* Even with a weaker backbone model, CC-RAG still generates a plausible causal explanation linking Andrew Yang's blockchain advocacy to increased awareness and adoption of Bitcoin. Unlike LLaMA with RAG, which fails to retrieve or generate relevant content about Yang, CC-RAG remains on-topic and structurally coherent. It shows that thematic scaffolding can preserve response quality even when retrieval is imperfect.
- **Weaknesses of CC-RAG:** *Redundant and Repetitive Reasoning:* The response suffers from verbosity and reiterates the same causal link multiple times. Compared to the more concise output in the full GPT-4o version, this ablated response lacks linguistic refinement and introduces some circular reasoning that may weaken its persuasive impact.

A qualitative analysis of BP Query 2, found in Table 11 shows the following:

- **Strengths of CC-RAG:** *Thematic Alignment and Self-Reinforcing Reasoning:* Despite the ablated model, CC-RAG generates a logically consistent and thematically relevant explanation. It correctly identifies a self-reinforcing cycle where economic instability boosts Bitcoin demand, which in turn drives up its price and perceived value. In contrast, LLaMA with RAG fails to ground its answer in the query and provides unrelated political context; Zero-Shot LLaMA does not respond meaningfully at all.
- **Weaknesses of CC-RAG:** *Repetitive Expressions:* The response reiterates the same causal loop multiple times, with slight rephrasings. Compared to the more concise and structured output from the full system, this version could benefit from tighter summarization.

Query	How does the accumulation of undigested glucocerebroside lead to splenomegaly and hepatomegaly in Gaucher disease?
CC-RAG	In Gaucher disease, the body is unable to produce the enzyme glucocerebrosidase properly, resulting in the accumulation of glucocerebroside fat. This buildup occurs in various organs, including the spleen and liver, causing them to enlarge, leading to splenomegaly and hepatomegaly and a swollen abdomen.
GPT-4o with RAG	In Gaucher disease, the accumulation of undigested glucocerebroside in cells occurs due to the hereditary deficiency of the enzyme glucocerebrosidase, which is responsible for breaking down glucocerebroside. This deficiency impairs the enzyme's activity, leading to the buildup of glucocerebroside, particularly in white blood cells like macrophages. These macrophages, filled with undigested glucocerebroside, are referred to as ""Gaucher cells."" The accumulation of Gaucher cells occurs in various organs, including the spleen and liver. The buildup of these lipid-laden cells leads to the enlargement of these organs—splenomegaly (enlargement of the spleen) and hepatomegaly (enlargement of the liver). The presence of these abnormal cells in high quantities interferes with the normal functioning of these organs and causes them to increase in size due to the storage of undigested material within the cells. This accumulation leads to the characteristic symptoms and complications associated with Gaucher disease.
Zero-Shot GPT-4o	In Gaucher disease, the accumulation of undigested glucocerebroside results from a deficiency in the enzyme glucocerebrosidase, which is necessary for its breakdown. This genetic disorder leads to the accumulation of glucocerebroside within lysosomes inside macrophages, which are called Gaucher cells when they contain these lipid-laden inclusions. The buildup of Gaucher cells primarily occurs in the spleen, liver, and bone marrow. In the spleen and liver, the accumulation of Gaucher cells causes these organs to become enlarged, leading to splenomegaly (enlarged spleen) and hepatomegaly (enlarged liver). The spleen and liver function in filtering and breaking down old blood cells and other cellular debris, so the infiltration and proliferation of Gaucher cells disrupt normal organ function, causing these organs to swell as they try to cope with the excessive storage material. Additionally, the accumulation impairs normal tissue architecture and may cause fibrosis and inflammation, further contributing to organ enlargement and dysfunction.

Table 8: Responses from different approaches for GD Query 1

Tables 12 and 13 show the responses from the CC-RAG and the baseline approaches to two sample queries from the BP dataset for the model substitution study.

A qualitative analysis of Model Substitution GD Query 1, found in Table 12 shows the following:

- **Strengths of CC-RAG:** *Causal Alignment with the Query:* CC-RAG correctly identifies the high-level causal chain—enzyme deficiency leads to substrate accumulation, which causes organ enlargement. It provides a coherent, top-down explanation aligned with the structure of the question, avoiding the excessive and fragmented repetition seen in LLaMA with RAG.

- **Weaknesses of CC-RAG:** *Redundancy:* The response repeats the same causal claim (accumulation leads to organ enlargement) using slightly different wording without introducing any new biological insight.

A qualitative analysis of Model Substitution GD Query 2, found in Table 13 shows the following:

- **Strengths of CC-RAG:** *Causal Alignment with the Query:* CC-RAG correctly identifies the high-level causal chain, enzyme deficiency leads to substrate accumulation, which causes organ enlargement. It provides a coherent, top-down explanation aligned with the structure of the question, avoiding the excessive and fragmented repetition seen in LLaMA with RAG.

- **Weaknesses of CC-RAG:** *Vague Terminology and Incorrect Wording:* The response includes phrases like “lipid-laden macrophags” and “Gauchen’s disease,” which, while suggestive, contain spelling errors.

Query	Why does glucocerebroside primarily accumulate in macrophages in Gaucher disease?
CC-RAG	In Gaucher disease, a mutation in the glucocerebrosidase (GCase) gene leads to reduced activity of the GCase enzyme, resulting in the accumulation of glucocerebroside. Macrophages, which are responsible for breaking down cellular debris, are particularly affected because they actively ingest and process large amounts of glucocerebroside-rich materials. The low GCase enzyme activity in these cells prevents proper degradation of glucocerebroside, leading to its accumulation primarily in macrophages.
GPT-4o with RAG	Glucocerebroside primarily accumulates in macrophages in Gaucher disease because the monocyte/macrophage lineage is preferentially altered due to their role in eliminating erythroid and leukocyte cells, which contain large amounts of glycosphingolipids. This cellular process leads to an accumulation of glucocerebroside (GlcCer) in these cells, initiating the characteristic features and complications of the disease.
Zero-Shot GPT-4o	Gaucher disease is a genetic disorder caused by a deficiency in the enzyme glucocerebrosidase. This enzyme is responsible for breaking down glucocerebroside, a type of glycolipid. In the absence or reduced activity of glucocerebrosidase, glucocerebroside accumulates within cells. Macrophages are part of the body's immune system and their function is to engulf and degrade various substances, including lipids, through a process called phagocytosis. In Gaucher disease, macrophages ingest cells or cellular debris containing glucocerebroside but are unable to degrade it due to the enzyme deficiency. As a result, the undegraded glucocerebroside accumulates within the lysosomes of macrophages, which then transform into Gaucher cells. These engorged macrophages are commonly found in the liver, spleen, bone marrow, and other organs. Therefore, glucocerebroside primarily accumulates in macrophages because these cells are key actors in clearing cellular debris and rely on glucocerebrosidase for the breakdown of ingested glucocerebroside. The deficiency of this enzyme leads to the pathological buildup of the lipid within macrophages.

Table 9: Responses from different approaches for GD Query 2

Query	How did Andrew Yang's advocacy for blockchain technology influence Bitcoin's adoption?
CC-RAG	Andrew Yang, a 2020 US presidential candidate, advocated for the adoption of blockchain technology. His advocacy was seen as a positive development for the cryptocurrency space. This increased interest in blockchain technology likely led to increased awareness and adoption of Bitcoin, a prominent cryptocurrency that utilizes blockchain. As a result, Bitcoin's user base and market capitalization grew. The increased adoption of cryptocurrencies, including Bitcoin, has led to greater mainstream recognition and acceptance of digital currencies. This shift in perception has contributed to the expansion of the cryptocurrency market as a whole. The growth of the market has attracted new investors and users, further increasing Bitcoin's popularity. Overall, Andrew Yang's advocacy for blockchains played a role in the increased adoption and recognition of Bitcoin.
LLaMA with RAG	There is no mention of Andrew Yang in the provided text. However, the text does mention that many of bitcoin's early fans, including libertarians, saw it (bitcoin) as a means to let ""individuals control their own financial destinities, free form the oversight"" of banks and government. The text also mentions that Satoshi Nakamato, the creator of Bitcoin, proposed the creation of Bitcoin as a ""way to make payment without relying"" on traditional banking systems. The text does not provide any information on how Trump's views on the subject have changed over time. It does mention Trump's pledge to ""create a strategic nation reserve for government's bitcoin holdings"" and to ease regulatory burden for the crypto-industry. However, it does not discuss how this has changed over the course of his presidency. It is worth noting that the text mentions that many crypto-investors are ""cheering Trump's promise"" to ease regulations and to create the national reserve. However it does mention the views of some individuals, including Tim Swansom, Joel Krugger, and James Slussar, who express skepticism about the potential of Bitcoin. They highlight its volatility and lack of widespread use as a payment method or store of wealth. Overall, the views on bitcoin expressed in the text are mixed, with some investors and analysts viewing it as having potential, while others express skepticism. There is no clear indication of how Trump views Bitcoin or how his views may have changed. However, the overall tone of the text suggests that many investors and some analysts are optimistic about the future of Bitcoin and the crypto market. This optimism is driven in part by Trump's pledges to ease regulation and create a national reserve, as well as the approval by the SEC of a series of exchange-tried funds that allow investors to buy Bitcoin directly. It is also driven by the growing mainstream adoption of Bitcoin by major financial institutions and asset managers. As the text notes, Bitcoin has seen a significant increase in mainstream adoption in recent months, with the likes of BlackRock and other global asset manangers entering the market. Overall, the view of the subject is that Trump's advocacy has likely had a positive impact on the adoption of blockchain technology and Bitcoin, at least in the short term. His pledges to create regulatory clarity and to support the development of a national Bitcoin.
Zero-Shot LLaMA	N/A ...

Table 10: Responses from different approaches for BP Query 1

Query	How did global economic uncertainty contribute to Bitcoin’s price rise?
CC-RAG	Global economic uncertainty leads people seeking alternative assets, as they perceive Bitcoin as such. This increased demand causes Bitcoin’s supply to be scarce, further driving up its price. As a result, Bitcoin becomes an attractive investment option, as its value rises in tandem with economic uncertainty, creating a self-reinforcing cycle. The increasing price of Bitcoin, in turn, reinforces its status as a reliable store of wealth, making it an even more attractive safe-harbor asset during times economic uncertainty and volatility. The resulting surge in demand for Bitcoin leads to a further increase in its value, solidifying its position as a sought-after alternative asset. This cycle of increasing demand and rising value has contributed to Bitcoin’s price rise during periods of global economic turmoil. As economic uncertainty persists, the price of bitcoin continues to rise, as more investors seek it as a secure and reliable investment option. The price of the cryptocurrency continues to appreciate as more people view it as an attractive alternative to other assets.
LLaMA with RAG	There is no information in the provided text about global economic uncertainties contributing to Bitcoin’s price rise. The text mentions that the "prospect of a favorable regulatory" environment and Donald Trump’s re-election contributed to Bitcoin hitting $100k. It also mentions that Bitcoin’s market movements can be influenced by "speculation that the U.S. will build a BTC Reserve". However, there is no mention of global economic instability or uncertainty. The text does mention the "sudden spectacular downfall" of the FTX exchange, which led to a 20% drop in Bitcoin’s value. However this is not presented as contributing to the price rise, but rather as a concern about the sustainability of the price increase. Therefore, the answer to this question cannot be determined from the provided information. However, one of the experts mentioned that there are concerns that the current price rise is a bubble and that the value of crypto, including Bitcoin, is based on speculation rather than inherent value, which is related to global economic stability. ...
Zero-Shot LLaMA	N/A. I don’t have information on the 2024 US presidential elections, or the subsequent price of Bitcoin.

Table 11: Responses from different approaches for BP Query 2

Query	How does the accumulation of undigested glucocerebroside lead to splenomegaly and hepatomegaly in Gaucher disease?
CC-RAG	Mutations in the GCase enzyme gene lead to low activity of the enzyme. This results in an accumulation of the glucoceroide. The buildup leads to enlargement of the spleen (splenomegly) and liver (hepatomaglgy).
LLaMA with RAG	The macro The macrophage that clears these cells is unable eliminate the product, leading to the cells turning into'Gaucher cell' and accumulate in the spleen and liver. The accumulation of these cells in the liver (hepatomegally) and spleens (splenomegly) causes them to become enlarged. This is due to the inability of the macrophagues to eliminate these cells, leading them to accumulate in these organs. Gaucher is characterized by enlargement of liver and the spleens. The accumulation leads to the enlargement of these organs and causes them to become swollen. It is due the inability to eliminate waste product from the body, which leads to accumulation in the organs. Gaucher disease is characterized the accumulation of undigest glucocerbroside in the cells, which causes the cells to turn into Gaucher. These cells accumulate in liver and spleen. They are unable to eliminate this waste product which leads them to turn in to Gaucher and accumulate in liver and splee. The cells that clear the cells are unable to do so, leading to the accumulation in the liver spleens and other organs. The macro phages that clear these cells are ina ble to eliminate the waste product, leading them turning into Gauchercells. Due to this, the cells accumulate the liver (heptomegely) and the splees (splomegley). This is due to inability to elminate the waste prdoduct from the body. In Gaucherdisease, cells are inaable to digest the glucoceresbrodie. As a result, they accumulate and turn into Gauchers. Thesecells accumulate in the splees and liver and cause them to become swollen. The inability to digest glucocorbrodie leads to accumulation and swelling of the splee and liver due to the inability of the macro phages to clear the wate product. Glucorbrode accumulates in cells and leads to Gauchaers. The inability leads ...
Zero-Shot LLaMA	"The accumulation of glucocèrebroside is toxic to the macrophages of the spleen and liver, causing them to swell and accumulate more glucocérébroside, leading to splenic and hepatic enlargement. N/A. The question is asking for a summary of how the accumulation leads to splenohepatomegally, not how it causes toxicity to macrophage cells. However, the information in the previous response is relevant to the question. Here is a revised summary. The toxic effect of the accumulation causes the macrophaege cells to swell, leading the splean and liver to enlarge. Splenomegly and hepatomagly result from the accumulation. The toxic effects of the accumulated glucocerbroside cause macrophagic cells to become engorged and swell, causing the splein and liver cells to enlarge, resulting in splenomgly and heptomegla. N/A. The question asks for a summery of how glucocorbroside leads to spleomeg and heptomgaly. A summary of the process is: The toxic effects cause macrophaeg cells to be engorgd and swell. The splein enlarges and the liver enlarges. Splenomgaly results from the spleins enlargement and hepatmogly results from liver enlargement. The accumulation of the glucorbroide is the cause of the toxic effects. This summary does not answer the question directly. It provides an indirect answer by describing the process in more detail. An answer that is more direct is: The accumulation causes macrophge cells in the sple and liver to become engorgded and swell and cause the sple to enlarge and the lver to enlarge resulting in spleomegly an heptomogly. Another answer is: Glucorbrose accumlation causes the sple and lver cells to engorgde and swell resulting in the enlargement of the organs and the development of splenmeagly and hepomagly. The toxic effect on macrophag cells in sple and lve causes the organs to enlarge in spleno and hepmagly, The accumulation leads the sple an lver cells to swel and become engored resulting in enlargement of sple and hepet and the deveolment of splenoagly an hepmogly, The accumulatoin of glucorbrise causes ...

Table 12: Responses from different approaches for Model Substitution GD Query 1

Query	Why does glucocerebroside primarily accumulate in macrophages in Gaucher disease?
CC-RAG	The mutation in the GCase enzyme leads to a deficiency in its activity, resulting in the buildup primarily in the body's macrophagic cells. The hallmark of Gauchen's disease is thus the accumulation and presence of these lipid-laden macrophags. The presence of this lipid accumulation in these cells leads to various manifestations of disease. Disease manifestations arise from the accumulation primarily in these macrophag cells.
LLaMA with RAG	"The accumulation of glycolipoprotein in macro phages is thought to result from the inability of the enzyme glucoc erebrosidase to degrade glucos ylceramide, which is a glycolipo protein. Gluco cerebroside is a type of glyco lipid that is primarily found in the lysosomes of macro phage cells. The accumulation occurs because the enzyme that breaks down glucos e ceramide is deficient or absent. This leads to the buildup of glucoc eremide in the macro ph age cells, resulting in the characteristic symptoms of Gauch er disease. Leave the summary as is. The accumulation of gluccerebroide primarily occurs in macrophas due to the inability to degrade glucceramide, a glycoprotein, resulting from a deficiency of the glucocerobidase enzyme. Gluccerebroside is a lipid primarily found within the lysosome of macrophas. The buildup of gluceceramide leads to symptoms characteristic of Gauche disease. The accumulation primarily occurs due to a lack of the GBA enzyme, leading the body to store glucccceramide. Glucecerbrosides primarily accumulate within macrophas because of the inability for the body break down gluce ceramide. Glucocerebroides primarily accumulate due to an inability to break down glucocermide. The enzyme responsible for this is deficient. Glucoresbrosdes primarily accumulate because the body cannot break down the glycoliprotein gluccermide due to deficiency in the enzyme. The body stores glucocerbosides primarily in macrophasis due to inability to metabolize glucec ceramide, leading symptoms of GD. Glucoresbroid primarily accumulates due to deficient GBA. The inability to breakdown gluce cermide leads to accumulation. Glucosecerbroid is primarily stored in macrophase due to lack of Gcase enzyme. Due to deficiency of GCA enzyme, glucoceresbroid accumulates in macropha cells. Gluecerboides primarily accumuate due to body's inability to digest glucecremid. Glcceresbros is primarily accumlated in macro phase due to GBA deficiency. Answer the question with the summary: Glucoceresbrode primarily accumlates due to enzyme deficiency, resulting the body storing the glyco protein in macro phases. Glcuecerebrode accumulates primarily in the body's macroph ...
Zero-Shot LLaMA	"N/A. I was unable to find any information about glucocorebroside. However, I can tell you that Gaucher's disease is a genetic disorder caused by a deficiency of the enzyme glucocerase, which is necessary for the breakdown of glucocorticoids. Glucocerebrosides primarily accumulate within the cells of the reticuloendothelial system, particularly in macrophasges. This accumulation is due to the inability of the deficient enzyme to break down the glucocercerobrosides, which then accumulate and cause the various symptoms of the disease. The accumulation of glucoscerobroside within macrophage cells leads to the formation of Gaucher cells, which are characteristic of the disorder. Gaucher cell are large, lipid-laden cells that are often found in the spleen, liver, and bone marrow. They are named after the French physician Philippe Gaucher, who first described the disease in 1882. Glucoscerbroside is a type of glycosphingolipid, a complex lipid that is composed of a sphingosine backbone and a carbohydrate group. It is an important component of cell membranes and plays a role in cell-to-cell communication and signaling. However, in the case of Gausher's disease, the accumulation of glucoscertbroside in macrophase cells leads to cellular dysfunction and disease. Glucocorebroside is not a known term in medical literature. However based on the information provided, it seems that the question is asking about glucoscebroside, which accumulates in macrophasis in Gauchers disease. The accumulation is primarily due to a deficiency in the enzyme glucocerobrasidase, which breaks down glucocertbroside into glucose and ceramide. The glucocerboside accumulates within the macrophasis and leads to cellular damage and dysfunction. The accumulated glucocorbroside within the macrophasis is often referred to as Gaucher cells. Gaucher s disease is an autosomal recessive genetic disorder that affects approximately 1 in 20,000 individuals. It is caused by mutations in the GBA1 gene, which encodes the enzyme glucose cerbrosidase. The disease is characterized by the accumulation of glucoserbroside in macrophysics, leading to cellular damage and dysfunction. Treatment of Gauches disease typically involves enzyme ...

Table 13: Responses from different approaches for Model Substitution GD Query 2

Figure 3: KG Representing Causal Relationships in Bitcoin Price Fluctuations