From Generating Answers to Building Explanations: Integrating Multi-Round RAG and Causal Modeling for Scientific QA

Victor Barres, Clifton McFate, Aditya Kalyanpur, Kailash Karthik Saravanakumar,
Lori Moon, Nati Seifu, Abraham Bautista-Castillo
Elemental Cognition Inc.
Correspondence: victor.barres@gmail.com, mcfateclifton@gmail.com

Abstract

Application of LLMs for complex causal question answering can be stymied by their opacity and propensity for hallucination. Although recent approaches such as Retrieval Augmented Generation and Chain of Thought prompting have improved reliability, we argue current approaches are insufficient and further fail to satisfy key criteria humans use to select and evaluate causal explanations. Inspired by findings from the social sciences, we present an implemented causal QA approach that combines iterative RAG with guidance from a formal model of causation. Our causal model is backed by the Cogent reasoning engine, allowing users to interactively perform counterfactual analysis and refine their answer. Our approach has been integrated into a deployed Collaborative Research Assistant (Cora) and we present a pilot evaluation in the life sciences domain.

1 Introduction

As Large Language Models (LLMs) demonstrate impressive performance on a wide variety of challenging tasks, there is intense interest in applying them to causal question-answering in complex domains such as life sciences. Examples of real queries asked in drug discovery research include:

- "How does epigenetic dysregulation of neurotrophins impact AD risk?"
- "What are the molecular pathways involved in the tumor environment of breast cancer?"

Questions like these, which we refer to as *complex causal questions*, are defined by several challenging characteristics. First, good answers are *causal and predictive*, requiring the resolution of causal factors to predict an unseen outcome. This resolution often requires *multi-step inference* as well as integrating information from *multiple sources*. Additionally, *multiple correct answers* arise from differing but consistent sets of assumptions.

Applying LLMs to problems with these characteristics can be stymied by the opacity of their decision making process and propensity for hallucination (Marcus, 2020). As such, there has been substantial effort to develop techniques that reduce hallucinations and equip LLMs with observable inferential steps such as Retrieval Augmented Generation (RAG) and Chain of Thought prompting (CoT) (Lewis et al., 2020; Wei et al., 2023). However, causal question answering remains particularly challenging (Bondarenko et al., 2022). We believe one reason is that prior research often neglects the processes by which humans select and evaluate causal explanations.

In this paper, we summarize criteria identified from a lengthy history of research in the social sciences as well as the shortcomings of existing LLM approaches (Miller, 2019). We then present a novel neuro-symbolic approach that addresses these shortcomings by using an executable causal model to guide iterative RAG. The resulting causal graph is backed by the *Cogent* Reasoning Engine, enabling interactive exploration of counterfactual scenarios. Our approach has been deployed for pilot users as a part of an existing life sciences research tool, *Cora* (Arsanjani and Brown, 2023). We evaluate performance on real queries from these pilot users.

2 Background

What makes an answer good or not depends on the task and context of its question. We begin by briefly summarizing findings from the social sciences that shed light on this topic for causal explanations and discuss where current LLM approaches fall short.

2.1 What Makes a Good Explanation?

Answers to complex causal questions have some obvious requirements: they must be coherent, relevant, and non-circular (Keil, 2006). Adding to

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 515–522
April 30, 2025 ©2025 Association for Computational Linguistics

these, we summarize the findings by Miller (2019) who suggest key criteria that guide selection and evaluation of explanatory answers.

First, explanations are generated and evaluated *selectively*, based on a *causal lens* reflecting preexisting biases and conceptual models (Miller, 2019). While there are potentially infinite framings for a given question, in general, Miller (2019) argue that good answers appeal to causal factors rather than probabilistic associations (see also Lombrozo (2006)). Bechtel and Abrahamsen (2005) highlight the central role of the notion of causal mechanism in scientific explanations in particular. Furthermore, they argue explanations are *contrastive* in that they are interpreted relative to an explicit or implicit foil (Miller, 2019).

Finally, explanations are *transactional* as they involve an attempt to communicate an understanding (Keil, 2006). Their causal framing is dependent on the expectations of the listener. Aligning on a conceptual lens is often interactive, making explanation generation a social process (Miller, 2019).

2.2 LLMs for Complex Causal QA

Retrieval Augmented Generation (RAG) decomposes LLM inference into a retrieval step over external resources (e.g. Wikipedia) and a generation step which produces output based on them (Lewis et al., 2020). RAG allows LLM applications to use information not stored within their parameters, resulting in answers more likely to be relevant and grounded in real world documents.

Zhu et al. (2021) review showed that such "retrieve and read" RAG approaches have demonstrated impressive performance in one-hop QA tasks. However, they still struggle in complex QA where coherent non-circular answers require threading inferences across documents. Going beyond iterative RAG (Qi et al., 2021), Trivedi et al. (2023) interleave RAG with chain of thought prompting (Wei et al., 2023) to answer multi-hop questions, which both improves performance and results in a trace of the inferential justification.

However, performance remains far from perfect and these approaches miss many of the key criteria for human explanations described above. While chat systems can answer successive questions, the lack of a consistent causal lens increases the risk of hallucination over multiple turns and leads to answers that lack the inter-connectivity and focus of human causal explanations.

3 Approach

These shortcomings influenced our approach to creating a causal QA system. It must answer the question by providing an *explanation structured by a coherent causal lens*, adjust to user expectations via *interactive feedback*, and allow *contrastive exploration*. For life sciences research, it must also *justify* its answer with relevant citations.

These criteria merge aspects best expressed symbolically (e.g structured inference) with others best handled by generative methods (e.g. Natural Language Generation and Information Extraction). For this reason, we designed a neuro-symbolic architecture in which a verbal explanation is generated from an interactive solution graph, as shown in Figure 1, whose semantics are grounded in a cognitively inspired causal formalism.

The graph allows the user to add, remove, and edit each node and edge. Each concept and relation in the graph is backed by a formal model defined in the *Cogent* reasoning engine (Chu-Carroll et al., 2024). Thus, as the user manipulates the graph, the effect on the target concepts is recomputed in real time, producing a final labeling which we use to update an evidenced natural language answer.

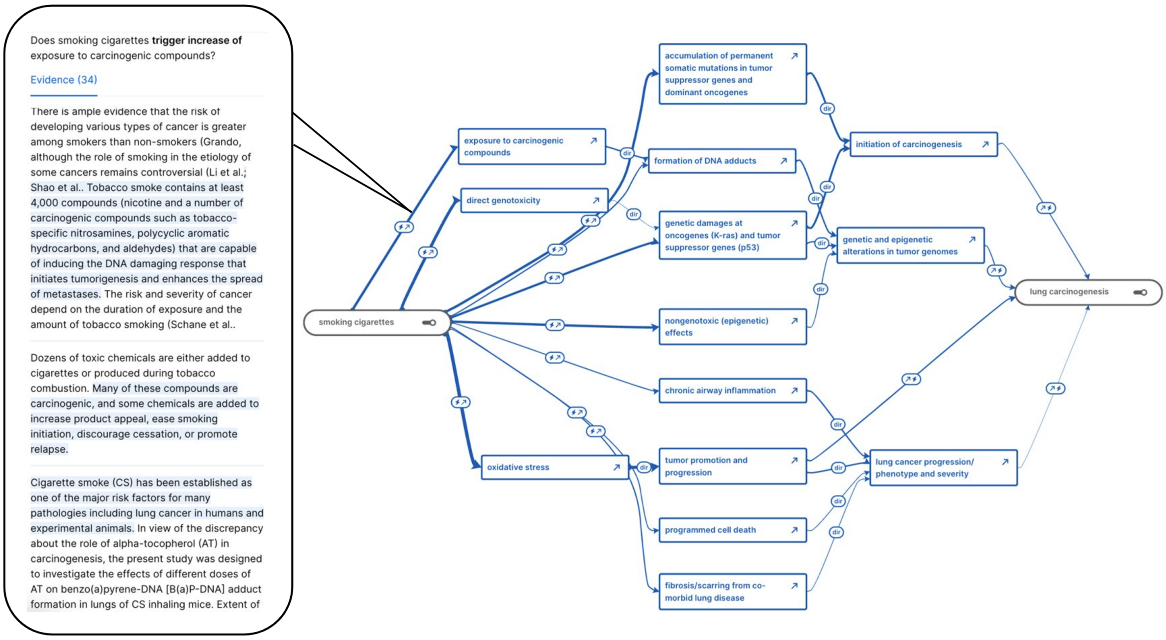
We begin by describing the solution graph and its underlying formal model. We then describe how that model acts as a scaffold for iterative RAG to construct the solution graph and NL answer.

3.1 Solution Graph

As discussed above, human explanations are selected and evaluated through restrictive causal lenses. To that end, we ground our search process and interface in a general causal model based on Qualitative Process Theory (QPT) (Forbus, 1984, 2019). In the following sections, we describe how QPT informs our solution graph and how it enables interactive reasoning. An instantiated example solution graph connecting smoking to lung carcinogenesis is shown in Figure 1.

3.1.1 Qualitative Process Theory

QPT is a formalism intended to capture how humans reason about continuous causal dynamics without precise numerical values. Under QPT, quantities are causally influenced by *processes*, and the effects of that influence propagate between quantities (Forbus, 1984, 2019). Approaches based on QPT have been used to annotate causal models in natural language (Friedman et al., 2022).


Does smoking cigarettes **trigger increase of** exposure to carcinogenic compounds?

Evidence (34)

There is ample evidence that the risk of developing various types of cancer is greater among smokers than non-smokers (Grando, although the role of smoking in the etiology of some cancers remains controversial (Li et al.; Shao et al.. Tobacco smoke contains at least 4,000 compounds (nicotine and a number of carcinogenic compounds such as tobacco-specific nitrosamines, polycyclic aromatic hydrocarbons, and aldehydes) that are capable of inducing the DNA damaging response that initiates tumorigenesis and enhances the spread of metastases. The risk and severity of cancer depend on the duration of exposure and the amount of tobacco smoking (Schane et al..

Dozens of toxic chemicals are either added to cigarettes or produced during tobacco combustion. Many of these compounds are carcinogenic, and some chemicals are added to increase product appeal, ease smoking initiation, discourage cessation, or promote relapse.

Cigarette smoke (CS) has been established as one of the major risk factors for many pathologies including lung cancer in humans and experimental animals. In view of the discrepancy about the role of alpha-tocopherol (AT) in carcinogenesis, the present study was designed to investigate the effects of different doses of AT on benzo(a)pyrene-DNA [B(a)P-DNA] adduct formation in lungs of CS inhaling mice. Extent of

smoking cigarettes
exposure to carcinogenic compounds
direct genotoxicity
accumulation of permanent somatic mutations in tumor suppressor genes and dominant oncogenes
formation of DNA adducts
genetic damages at oncogenes (K-ras) and tumor suppressor genes (p53)
nongenotoxic (epigenetic) effects
chronic airway inflammation
oxidative stress
tumor promotion and progression
programmed cell death
fibrosis/scarring from co-morbid lung disease
initiation of carcinogenesis
genetic and epigenetic alterations in tumor genomes
lung cancer progression/ phenotype and severity
lung carcinogenesis
dir

Figure 1: Example solution graph connecting smoking and lung cancer. Quantity nodes are blue if they have positive polarity and red otherwise. State nodes are grey, with a toggle indicating whether the are active or not. Users can view evidence for each edge, manually add or remove nodes and edges, and perform contrastive analysis by manipulating node polarity. On the left is evidence for the initial edge from smoking to carcinogen exposure.

Like prior work, we take inspiration from QPT's influence mechanism, but we expand our approach to include `States` and a corresponding `Triggers` causal relationship. In life sciences, `Quantities` encompass fluents like *blood pressure*, while `States` represent booleans or specific fluent values such as *having diabetes* or *high blood pressure*. In our solution graph, quantities and states are nodes. Quantities can be one of *increasing*, *decreasing*, or *stable*. States can be either *active* or *inactive*.

In Figure 1, the initial state (smoking cigarettes) is active. It triggers increases in downstream quantities (e.g. oxidative stress). Each edge in the solution graph is either an `Influences` or a `Triggers` relation. Influences hold between two quantities and are either direct or inverse. For instance, in life sciences, an *increase in medication dosage* might inversely influence (decrease) *symptom severity*.

`Triggers` define causal relationships involving states, allowing them to act as tipping points for quantity changes. For example, the *detection of foreign pathogens* (a `State`) might trigger an *increase in white blood cells*.

3.1.2 Interactive Graph Reasoning

The solution graph is backed by a formal model defined in the *Cogent* reasoning engine (Chu-Carroll et al., 2024). Cogent is a commercial multi-heuristic reasoning engine built on Gebser et al. (2012)'s *clasp* answer set programming solver. Cogent executes models written in a constrained English language with broad semantics that supports term definitions, rules, (hard/soft) constraints and objective functions (Chu-Carroll et al., 2024). Cogent propagates known values (e.g. increasing/decreasing) through the graph and outputs a complete labeling for quantities and states.

3.2 Iterative Graph Building

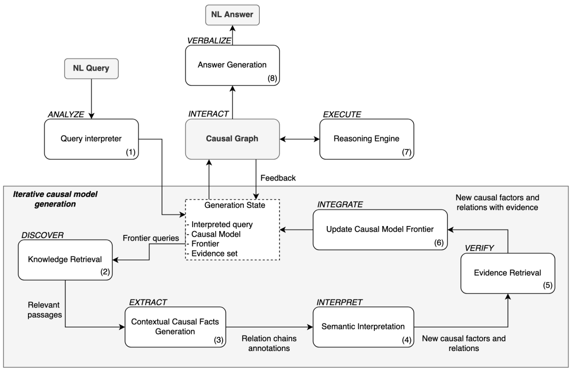


Figure 2: Examples of text annotated with proto-roles and the resulting solution graph relation

The solution graph is built incrementally using a forward-backward graph expansion approach based on A* search (Hart et al., 1968). Given a question, we begin by extracting independent and dependent variables as initial graph nodes (*Analyze* step). With these nodes as initial frontiers, graph expansion proceeds in a loop as shown in Figure 2 and explained below.

Although the approach can be used with any IR system, in this paper we present pilot results from integrating with an existing life sciences research tool, *Cora*, which processes and indexes PubMed documents with extracted domain concepts and document embeddings (Kalyanpur et al., 2024).

1. **Analyze** Prompt an LLM to extract independent and dependent entities from the user query. The goal is to understand how the independent entities (sources) control the behavior of the dependent entities (targets). These become the initial graph frontiers. This steps also allows the system to abstain from answering questions that are do not call for causal explanations.

2. **Discover** Query Cora for (a) documents relevant to understanding how sources causally affect targets and (b) documents addressing the causal effects of the sources or (c) the possible causes behind observation of the targets.

3. **Extract** Prompt an LLM to generate causal chain annotations using the retrieved documents, the QPT annotation format, and the current state of the causal graph. We require all chains to provide a full causal path from source to target.

   Initial attempts to generate influence and triggers relationships directly, as well as casual chains with unstructured source and target entities, struggled to produce precise and distinct chains. The result was often overlapping paths with near-synonymous nodes. One possible reason comes from the flexible nature of agent and patient argument selection in English verbs. This flexibility lead Dowty (1991) to deconstruct these classic semantic roles into collections of "proto-role" properties.

   Inspired by this work, we decompose our concepts and relations into combinations of "change" (quantities) and "value" (states) properties. The LLM is prompted to find causal relations between entities with these modifiers, which enforces a consistent framing for interpreting agents and patients in causal statements. Figure 3 contains example sentences, proto-role annotations, and the resulting solution graph nodes and edges.

4. **Interpret** As shown in Figure 3, each combination of attributes and causal relation corresponds to an edge between two nodes in our causal graph. We deterministically map each annotation to its Cogent QP concepts (quantity/state) relationships (influence/triggers).

5. **Verify** Given the new concepts and causal relationships generated, query Cora to retrieve evidence supporting each claim. Then, prompt the LLM to further refine selected evidence by extracting supporting passages. Relations lacking evidence are pruned, and remaining supported relations are advanced to the integrate step.

6. **Integrate** Extend the graph forward from the source frontier and backwards from the target frontier using the causal relations. At this point, the partial graph is amenable to user modification. Any remaining disconnected nodes become frontiers for the subsequent iteration: repeat the Discover, Extract, Interpret, Verify and Integrate steps.

3.3 Answer Generation

Cogent computes a labeling from the completed graph which is given, along with the graph and evidence, to an LLM for answer generation. Each statement in the answer derives from a causal path in the solution graph, citing evidence along that path. Thus, the rhetorical structure reflects the underlying causal model.

4 Evaluation: Life Sciences

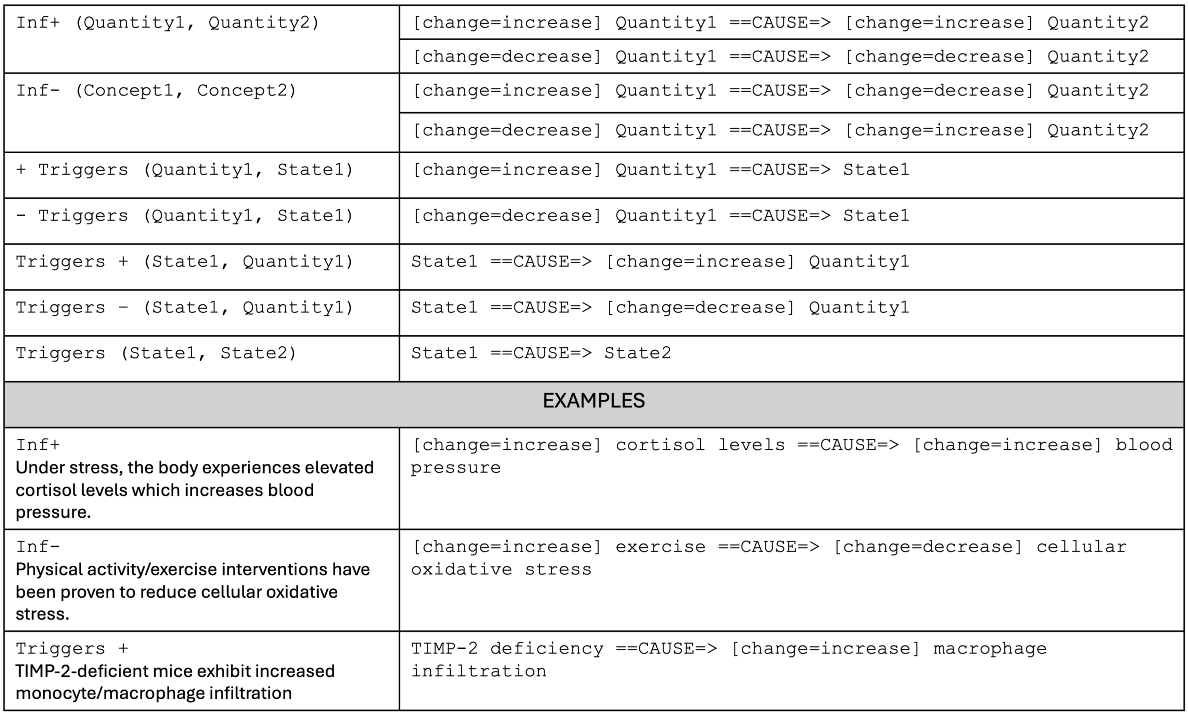
We report the results of an evaluation based on a set of 25 *multi-hop causal queries* sampled from pilot life sciences researchers using Cora in production. We compare the natural language answer generated by our approach to those from three commercially available services: GPT4-*Turbo*[1] (state of the art LLM) , *Perplexity*[2] (Commercial RAG using websearch), *Elicit*[3] (Commercial RAG using Semantic Scholar), and *Our solution*.[4]

[1] openai.com

[2] https://www.perplexity.ai/

[3] https://elicit.com/

[4] Answers generated without interactive user feedback.



Inf+ (Quantity1, Quantity2)	[change=increase] Quantity1 ==CAUSE=> [change=increase] Quantity2
	[change=decrease] Quantity1 ==CAUSE=> [change=decrease] Quantity2
Inf- (Concept1, Concept2)	[change=increase] Quantity1 ==CAUSE=> [change=decrease] Quantity2
	[change=decrease] Quantity1 ==CAUSE=> [change=increase] Quantity2
+ Triggers (Quantity1, State1)	[change=increase] Quantity1 ==CAUSE=> State1
- Triggers (Quantity1, State1)	[change=decrease] Quantity1 ==CAUSE=> State1
Triggers + (State1, Quantity1)	State1 ==CAUSE=> [change=increase] Quantity1
Triggers - (State1, Quantity1)	State1 ==CAUSE=> [change=decrease] Quantity1
Triggers (State1, State2)	State1 ==CAUSE=> State2
EXAMPLES	
Inf+ Under stress, the body experiences elevated cortisol levels which increases blood pressure.	[change=increase] cortisol levels ==CAUSE=> [change=increase] blood pressure
Inf- Physical activity/exercise interventions have been proven to reduce cellular oxidative stress.	[change=increase] exercise ==CAUSE=> [change=decrease] cellular oxidative stress
Triggers + TIMP-2-deficient mice exhibit increased monocyte/macrophage infiltration	TIMP-2 deficiency ==CAUSE=> [change=increase] macrophage infiltration

Figure 3: Decomposition of solution graph relations into proto-roles and examples of text along with proto-role annotations and the resulting graph relation

This dataset was curated to include answers that required multi-hop inference. In order to avoid confounds due to surface form variations and to facilitate the evaluation, the queries of our dataset were uniformly reformatted using the construction "How does X impact Y?".

4.1 Methodology

Each system was given each query and prompted to produce an answer with supporting/refuting evidence and cited sources. Our system implementation uses GPT4 for each of the prompted LLM calls. The approach requires no fine-tuned model, making it highly adaptable to new domains and opening avenues for reductions in speed and cost via fine-tuning.

Since each question could have multiple correct answers, our evaluation focuses on validity, verifiability, and relevance rather than a comparison to a single gold standard. To assess these characteristics, we designed the following rubric and had domain experts review each systems' results.

1. **Claim Density**: Average number of claims per answer. *A measure of the quantity of information provided.* (CLM Density)

2. **Citation Density**: Average number of real citations per claim. (CT Density)

3. **Source Hallucination Rate**: Percentage of citations that are not valid (real) scholarly sources. (HL Rate)

4. **Citation Rate**: Percentage of claims in the answer that are accompanied by real citations. (CT Rate)

5. **Justification Rate**: Percentage of claims that are a correct paraphrase of a real citation. *A measure of interpretation quality*. Claims with non-existent sources are unjustified. Since verification requires manual effort, we imposed a 5-minute time-limit for the domain expert to verify each claim. (JT Rate)

6. **Relevance Rate**. Percentage of claims that are justified and relevant to answering the question. (REL Rate)

Note that the measures from 4-6 get progressively stricter, as a justified claim must also be cited, and a relevant claim must also be justified. We also asked a domain expert to quantify the complexity of the explanation generated, recording:

1. **Maximum Number of Hops**: Maximum number of hops (relations) tying the source (X) to the target (Y) in a reasoning chain.

2. **Number of Concepts**: Number of concepts presented in the answer that are directly relevant to the explaining the mechanism.

4.2 Results

Our approach outperforms the comparison systems across all evaluated categories except for citation density, in which Elicit has a narrow advantage.

Beginning with our first 3 measures in Table 1, our solution beats competitors in *Claim Density* which measures the quantity of information presented in the answer. Looking at each claim's citations Ours, Elicit and Perplexity all reliably cite articles that exist (HL Rate) while GPT-4 has a high rate of hallucination. Perplexity, however, cites fewer articles for fewer claims, as evidenced by low *CT Density*.

System	CLM Density	CT Density	HL Rate
GPT4-Turbo	4.16	1.01	31.4%
Perplexity	4.76	0.59	0.01%
Elicit	5.00	**1.36**	0.01%
Our System	**5.36**	1.14	**0.00%**

Table 1: Multi-hop Query Results Measures 1-3

Evaluation measures 4-6 in Table 2 measure the supportability and quality of claims. Our system has the highest rate of cited, justified claims. The *Relevance Rate* is a more subjective measure of usefulness by our experts, obtained by considering how many justified claims in an answer they also label as relevant. Results show that our system outperforms the next best tool by nearly *26%*.

System	CT Rate	JT Rate	REL Rate
GPT4-Turbo	64.42%	27.88%	22.12%
Perplexity	32.77%	17.65%	11.76%
Elicit	98.40%	86.40%	60.80%
Our System	**98.51%**	**90.30%**	**86.57%**

Table 2: Multi-hop Query Results Measures 4-6

The answer complexity analysis shown in Table 3 adds another dimension to the results. A pure LLM solution such as GPT-4 Turbo generates answers with a high number of concepts and the longest reasoning chains. However, as shown in Table 1, most of its claims are unjustified and/or irrelevant. Elicit has a higher rate of justification and relevance but produces fewer concepts with fewer hops. Our system's answers combine high coverage and depth with justified relevant claims.

System	Max Hops	Number of Concepts
GPT4-Turbo	2.5 ±2.1	5.1 ±3.1
Perplexity	1.5 ±1.2	4.0 ±3.3
Elicit	0.8 ±0.6	3.3 ±3.2
Our System	2.1 ±0.7	7.5 ±2.4

Table 3: Multi-hop Query. Answer Complexity

4.3 Example: Multi-hop answer comparison

We conclude our evaluation with an illustrative comparison of the two best performing systems, Ours and Elicit. The answers are generated by both systems for the query "How does epigenetic dysregulation of neurotrophins impact AD (Alzheimer's Disease) risk?" shown in Table 4. At first glance, Elicit may provide a satisfying answer. However, analysis by a domain expert reveals:

- **Concept Granularity Mismatch** Whereas the question asked about "epigenetic dysregulation of neurotrophins" specifically, Elicit's answer opens (E1) with the overly broad phenomenon of epigenetic dysregulation. The "particularly" clause adds generic processes, not those specific to neurotrophin dysregulation. This contrasts with our system's opening statement (C1) which targets the right concept and further specifies specific processes.
- **Unjustified or Irrelevant Claims** A domain expert concluded that the references cited in E2, E3, and E4 do not back the associated claims. Conversely, all evidences cited by our system supported their associated claim. Elicit's last statement (E4) about therapeutic strategies is out of scope.
- **Vagueness and Lack of Depth** Most statements in the Elicit answer are vague and attempt to link source and target concepts directly. For example, E3 simply indicates "a link to AD risk" without expanding further. In contrast, our approach generates specific claims (e.g. a "decrease synthesis of BDNF") which appear in complex multi-hop chains.

5 Conclusions and Future Work

In this paper we have presented a novel approach to complex causal QA, implemented and evaluated as part of a life sciences Collaborative Research Assistant (Cora). We argue existing LLM approaches are insufficiently performant and fail to address key criteria by which humans generate and evaluate explanations. Drawing from the social sciences, we

	Query: How does epigenetic dysregulation of neurotrophins impact AD risk?
	Our System's Answer
C1	Epigenetic dysregulation of neurotrophins, particularly through increased promoter CpG methylation, leads to a down-regulation of brain-derived neurotrophic factor (BDNF) expression, which is associated with decreased synthesis of BDNF in neurons [1][2].
C2	This dysregulation is linked to reduced neuroprotection and has been implicated in the emergence of various neurological and psychiatric diseases [3][4].
C3	BDNF is crucial for cognitive function, promoting neurogenesis, neuronal survival, axonal and dendritic growth, synaptic plasticity, and maintenance in the central nervous system [5].
C4	Consequently, decreased levels of BDNF result in the reduction of synaptic proteins, leading to synaptic connection loss, impaired neuronal survival, and adaptability, which are critical factors in neurodegeneration [6].
C5	This pathway of dysregulation contributes significantly to Alzheimer's disease (AD) neuropathology, particularly through the increase of hippocampal neurofibrillary tangles (NFTs) and plaques (NPs), which are strongly correlated with cognitive decline and are hallmark features of AD [7][8].
C6	Thus, epigenetic dysregulation of neurotrophins, by decreasing BDNF expression and signaling, directly impacts AD risk by promoting mechanisms that lead to neuronal and synaptic degeneration.
	Elicit's Answer
E1	Epigenetic dysregulation, particularly in the form of DNA hypermethylation and histone deacetylation, has been implicated in the pathophysiology of Alzheimer's disease (AD) (Chouliaras 2010, Sanchez-Mut 2015, Millan 2014).
E2	This dysregulation can lead to a repressed chromatin state and reduced plasticity, potentially contributing to the onset and progression of AD (Sanchez-Mut 2015).
E3	The dysregulation of neurotrophins, a family of proteins that support the survival and growth of neurons, is a key aspect of this epigenetic dysregulation (Zusso 2018). In particular, the dysregulation of brain-derived neurotrophic factor (BDNF) has been linked to AD risk (Nativio 2018).
E4	This suggests that targeting the epigenetic dysregulation of neurotrophins, including BDNF, could be a potential therapeutic strategy for AD (Lardenoije 2015, Qureshi 2011, Daniilidou 2011).

Table 4: Comparison of Elicit's and Our system's answers to the query "How does epigenetic dysregulation of neurotrophins impact AD risk?"

designed our approach around an executable causal model which guides iterative RAG and grounds an interactive solution graph. Using real queries from pilot life sciences users, we demonstrate that our approach provides broader, deeper, and better evidenced answers than existing commercial systems.

In future work, we plan to expand causal frameworks to include alternatives to QPT. Ross (2021), for example, argue that life science research also uses a "pathway" model of causation that differs from a mechanistic view. We would like to allow users to design and align their own causal formalism to the solution graph. We also plan to extend our approach to include refuting evidence to counteract confirmation bias and identify competing causal theories.

References

Ali Arsanjani and Eric Brown. 2023. Built-with google ai: Reliable and transparent ai from elemental cognition. Avalaible at `https://shorturl.at/JvUWx`.

William Bechtel and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):421–441. Publisher: Elsevier.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. CausalQA: A Benchmark for Causal Question Answering. In *29th International Conference on Computational Linguistics (COLING 2022)*, pages 3296–3308. International Committee on Computational Linguistics.

Jennifer Chu-Carroll, Andrew Beck, Greg Burnham, David OS Melville, David Nachman, A Erdem Özcan, and David Ferrucci. 2024. Beyond llms: Advancing the landscape of complex reasoning. *arXiv preprint arXiv:2402.08064*.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Kenneth D Forbus. 1984. Qualitative process theory. *Artificial intelligence*, 24(1-3):85–168.

Kenneth D Forbus. 2019. *Qualitative representations: How people reason and learn about the continuous world*. MIT Press.

Scott Friedman, Ian Magnusson, Vasanth Sarathy, and Sonja Schmer-Galunder. 2022. From unstructured text to causal knowledge graphs: A transformer-based approach. *arXiv preprint arXiv:2202.11768*.

Martin Gebser, Benjamin Kaufmann, and Torsten Schaub. 2012. Conflict-driven answer set solving: From theory to practice. *Artificial Intelligence*, 187:52–89.

Peter Hart, Nils Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Aditya Kalyanpur, Kailash Saravanakumar, Victor Barres, CJ McFate, Lori Moon, Nati Seifu, Maksim Eremeev, Jose Barrera, Eric Brown, and David Ferrucci. 2024. Multi-step knowledge retrieval and inference over unstructured data. *Preprint*, arXiv:2406.17987.

Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology*, 57(Volume 57, 2006):227–254. Publisher: Annual Reviews.

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.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Gary Marcus. 2020. The next decade in ai: Four steps towards robust artificial intelligence. *Preprint*, arXiv:2002.06177.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Peng Qi, Haejun Lee, Oghenetegiri "TG" Sido, and Christopher D. Manning. 2021. Answering Open-Domain Questions of Varying Reasoning Steps from Text. *arXiv preprint*. ArXiv:2010.12527 [cs].

Lauren N. Ross. 2021. Causal Concepts in Biology: How Pathways Differ from Mechanisms and Why It Matters. *The British Journal for the Philosophy of Science*, 72(1):131–158. Publisher: The University of Chicago Press.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. *arXiv preprint*. ArXiv:2212.10509 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. *arXiv preprint*. ArXiv:2101.00774 [cs].