PathwayLM: Multihop Mechanistic Pathways for Biomedical Language Model Reasoning

Anonymous ACL submission

Abstract

We introduce a high-throughput framework to semi-automatically construct multihop reasoning datasets in the biomedical domain. We use a neuro-symbolic information extraction (IE) system to extract individual biomedical interactions, followed by a constraint-based path construction algorithm that aggregates complete paths and filters out noise. We use this framework to construct over 5 million semantically consistent 2-hop paths from 4M biomedical publications. We also manually curate 137 paths into a "gold" test partition. We use this dataset to evaluate the capacity of LLMs to mechanistically reason in the biomedical domain. Our evaluation shows that: (a) biomedical reasoning remains an open research problem; and (b) a promising practical avenue that doubles reasoning performance is to use the IE system as scaffolding for LLM reasoning.

1 Introduction

The volume of biomedical literature is expanding exponentially (e.g., PubMed[1] has indexed more than 1M publications per year in the past decade), making it impossible for researchers to manually track all potential mechanistic explanations of diseases. While information extraction (IE) systems can identify individual biomedical interactions (Kim et al., 2009; Islamaj et al., 2024) (e.g., *Protein A phosphorylates Protein B*), true biological insight requires *multihop reasoning* that connects disparate biomedical interactions from different publications to form a plausible mechanistic hypothesis.

Current datasets for biomedical Question Answering (QA) and reasoning often suffer from two limitations: (1) they are manually curated, which limits their scale and update frequency, or (2) they focus on factoid retrieval rather than structured mechanistic chains (see Related Work).

Here, we introduce a scalable, automated method for generating multihop mechanistic paths at scale. Our approach uses a neuro-symbolic IE system[2] (Valenzuela-Escárcega et al., 2018) to extract individual biomedical interactions, filters them for quality, and assembles them into multihop reasoning paths. Some of these paths have been manually curated into a smaller "gold" test partition. The proposed framework: (a) is scalable: as new literature is published, the dataset can be automatically updated; (b) supports n-hop reasoning: while this paper focuses on 2-hop paths, the logic extends naturally to n-hop paths; and (c) is adaptable: our filtering constraints can be tuned to prioritize high precision (for gold-standard creation) or high recall (for exploratory hypothesis generation).

Using this dataset, we evaluate the capacity of 11 LLMs to reason in the biomedical domain, when provided with a variety of textual contexts (from multiple paragraphs to individual sentences containing interactions of interest). Our evaluation highlights that all LLMs perform poorly when provided with too much context. However, performance approximately doubles for a neuro-symbolic strategy, in which when the LLM uses as context only sentences containing interactions of interest with marked up evidence by the IE system.

Our contributions are as follows:

- We describe a high-throughput framework to semi-automatically construct multihop reasoning datasets in the biomedical domain. We couple a neuro-symbolic IE system with a constraint-based path construction algorithm that filters out noise. We use this framework to construct over 5 million semantically consistent 2-hop paths from 4M biomedical publications. We also manually curate 137 paths

[1] https://pubmed.ncbi.nlm.nih.gov

[2] REACH (REading and Assembling Contextual and Holistic mechanisms from text)

into a "gold" test partition.[3]

- We use this dataset to evaluate the capacity of LLMs to mechanistically reason in the biomedical domain. Our evaluation shows that: (a) despite the progress in reasoning algorithms, biomedical reasoning remains an open research problem; and (b) a promising practical avenue that doubles reasoning performance is to use a neuro-symbolic IE system as scaffolding for LLM reasoning.

2 Related Work

Reasoning with LLMs. Prompting strategies have been shown to elicit multi-step reasoning in large language models, most notably through chain-of-thought (CoT) prompting, which improves performance on arithmetic, symbolic, and commonsense reasoning tasks (Wei et al., 2022). Similar gains can be achieved in zero-shot settings, suggesting that reasoning abilities are latent in sufficiently large models (Kojima et al., 2022). However, single-path greedy decoding is often brittle. Methods such as self-consistency address this limitation by sampling and aggregating multiple reasoning trajectories (Wang et al., 2023), while least-to-most prompting improves generalization by decomposing complex problems into simpler subproblems solved sequentially (Zhou et al., 2023).

Subsequent work introduces additional structure through search, computation, or interaction. Program-of-Thoughts separates natural language reasoning from execution by generating runnable programs (Chen et al., 2023), Tree-of-Thoughts expands linear reasoning into a branching search process (Yao et al., 2023a), and ReAct interleaves reasoning with tool use to ground generation and reduce hallucinations (Yao et al., 2023b). While effective, these approaches are largely evaluated in open-domain settings. In contrast, our work targets reasoning within the complex domain of biological mechanistic pathways.

LLMs for the biomedical domain. Progress in biomedical language modeling has primarily come from domain-specific pretraining and task-focused fine-tuning. Training models on large biomedical corpora yields strong performance across core biomedical NLP tasks (Gu et al., 2021), with generative models such as BioGPT enabling end-to-end biomedical text generation and mining (Luo et al., 2022b). Scaling domain-specific data further benefits clinical applications, as shown by large EHR-trained models like GatorTron (Yang et al., 2022), while smaller models trained exclusively on curated biomedical text remain competitive on biomedical QA benchmarks (Bolton et al., 2024). Instruction-tuned medical LLMs extend these advances to question answering and dialogue, with systems such as Med-PaLM and open alternatives demonstrating strong task performance alongside persistent reliability challenges (Singhal et al., 2023, 2025; Toma et al., 2023; Zhang et al., 2023).

More recent work shifts attention from broad biomedical task performance toward explicitly modeling medical reasoning. Approaches that disentangle knowledge recall from reasoning improve interpretability and robustness (Jin et al., 2025), while clinically grounded methods align generation with physician-like reasoning pathways to reduce hallucinations (Wu et al., 2024). Efficient test-time adaptation further improves medical reasoning without full fine-tuning (Shi et al., 2024). While most of the work in biomedical reasoning with LLMs focuses on clinical and medical domains, our work focuses on using LLMs for *mechanistic reasoning* over biochemical interactions to discover consistent signaling pathways.

Biomedical relation datasets. Benchmarks like BioRED (Luo et al., 2022a) focus on single-hop relation extraction from localized text, lacking the multi-hop chains required for mechanistic understanding. Our dataset addresses this gap by encoding gold reasoning paths and introducing multi-level contextual granularity, enabling the systematic evaluation of compositional logic and explainability in biomedical models.

3 Dataset Construction

Our pipeline consists of three phases: (1) interaction extraction, (2) quality filtering, and (3) multi-hop path construction.

3.1 Large-Scale Interaction Extraction

A total of 4.19 million research articles from PubMed are initially considered for the corpus. After applying a series of filtering criteria, the corpus is reduced to 357,161 articles. Articles are retained if they satisfy at least one of the following conditions: (1) the citation count among the 4.19 million research articles exceeded the threshold defined by $2 + \lfloor 0.1 \times (2025 - \text{publication\_year}) \rfloor$; or (2)

[3]We release the dataset to facilitate research into symbolic explainability and large-scale biomedical reasoning.

the journal in which the article is published has an impact factor of at least 2. Furthermore, articles with non-empty textual content extracted by the REACH are only included in the final corpus. REACH operates by identifying: (a) entities: genes, proteins, chemicals, diseases, cells, organs, tissues, and biological processes, normalized to standard ontologies (e.g., *uniprot*, *go*, *fplx*, etc.); (b) events: mechanistic interactions such as phosphorylation, activation, or inhibition; (c) polarity: the type of the interaction, such as promotion or inhibition. After this extraction, we obtain 1.9 million individual interactions. Subsequently, the interactions are grouped according to their frequency, defined as the number of distinct articles providing evidence for each interaction. From these groups, only interactions supported by exactly two independent articles are selected. These steps substantially reduced the noise, retaining only those interactions with a high probability of representing valid biological facts. Applying this criterion yields an initial pool of 142,916 raw interactions.

3.2 Noise Filtering and Quality Control

Raw extractions contain noise due to parser errors or ambiguous sentence structures. To ensure the interactions were suitable for deductive reasoning and biological relevance, we applied strict filtering constraints: (1) the controller and controlled entities are required to be distinct; (2) interactions involving pronouns as entities are removed; and (3) interactions are retained only if both entities are categorized as bioprocesses (*mesh*, *go*, *frailty*, *bioprocess*) or if both entities contain an **uppercase** letter or a digit (example shown in Figure 3). After applying all these filtering criteria, the dataset is reduced to 107,263 interactions.

We expand the REACH-assigned entity labels using `spaCy` (`en_core_web_sm`) (Honnibal et al., 2020) to obtain more accurate and contextually grounded entity descriptions. We then performed an additional round of stop-word and pronoun filtering on the expanded descriptions. After these processing steps, the dataset contained a total of 101,097 interactions.

3.3 Multihop Path Construction

We partitioned filtered interactions into training, validation, and test sets, then linked interactions within each split to form multi-hop paths while ensuring strict disjointness (Table 2). To maintain consistency, we filtered out paths containing internal bioprocess entities or negated triggers. Construction statistics and final path counts are detailed in Tables 1 and 2.

4 Evaluation

To construct a high-quality gold standard for evaluation, we randomly sampled 320 2-hop paths from the test partition, ensuring all instances contained unique entity pairs and interactions. Each path underwent double-blind review by two independent annotators (from a pool of four), yielding substantial inter-annotator agreement (Cohen's κ 0.72). We filtered for strict unanimity, retaining only the 137 paths where both annotators agreed on correctness to serve as the ground truth.

For the experimental benchmark, we evaluated 11 LLMs (7 reasoning, 4 non-reasoning) using four query types: *Relation Existence*, *N-Hop Count*, *Intermediate Entity*, and *Polarity*. We assessed performance across five context levels of increasing noise: target sentences (with and without symbolic markup), sentence windows (left-target-right), paragraphs, and paragraph windows (see Appendix C for details). All models were restricted to a maximum output length of 512 tokens, with a limit of 10 retry attempts per query to mitigate format-based generation failures.

5 Results

Figure 1 demonstrates that performance peaks when input is restricted to single sentences with explicit symbolic markup. Removing markup or expanding context to paragraphs consistently degrades accuracy, suggesting that additional unstructured text acts as noise rather than signal. Reasoning models consistently outperform non-reasoning baselines, exhibiting more graceful degradation as context grows, whereas non-reasoning models often collapse beyond sentence-level inputs (Table 3). While accuracy generally declines as context increases, several models violate this monotonic trend by recovering performance on multi-paragraph inputs, suggesting that redundant evidence can partially offset the effects of context dilution. Among query types, *N-Hop* reasoning remains the most challenging, while *Polarity* prediction is relatively stable (Figure 2), likely due to reliance on local lexical cues. These results indicate that robust biomedical reasoning requires structured, interaction-centric evidence rather than simply larger context windows.

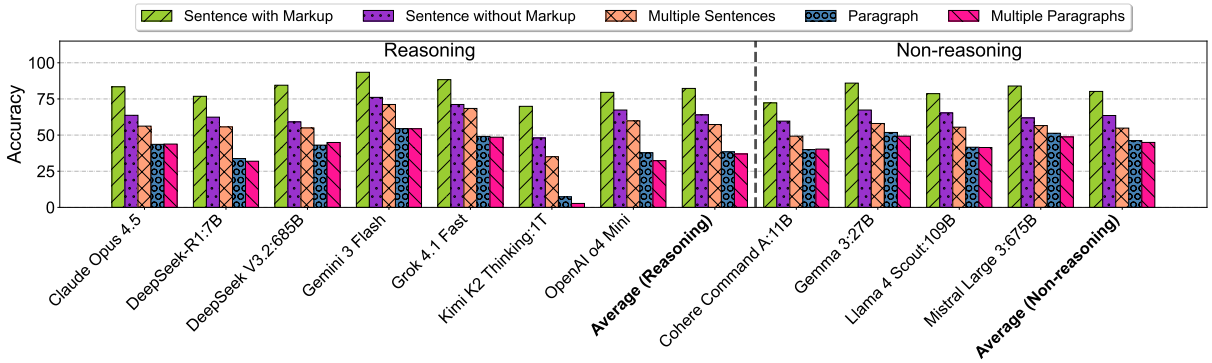

Figure 1: Impact of contextual complexity on model accuracy (aggregated across all query types). Results show consistent performance degradation as inputs expand from symbolic sentences to unstructured long evidence, such as multiple paragraphs.

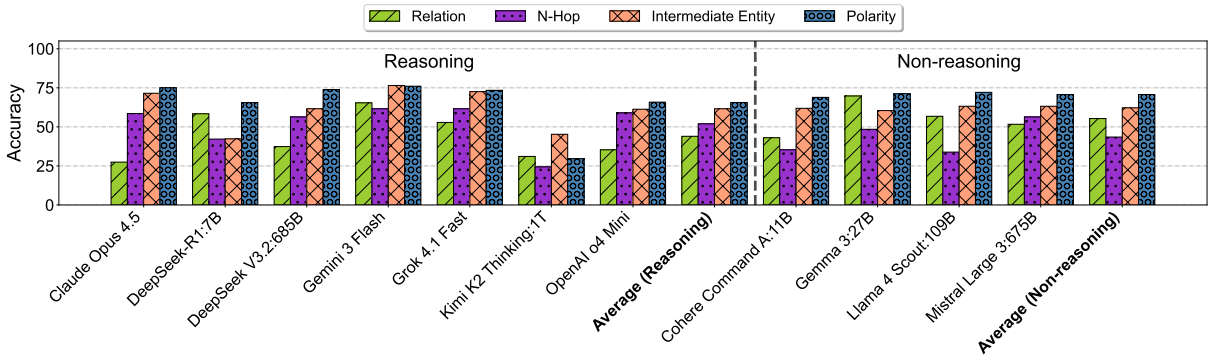

Figure 2: Performance breakdown by query type. Averaging across context types reveals *N-Hop* reasoning remains the consistent bottleneck, whereas robust *Polarity* prediction suggests reliance on local lexical cues rather than complex multistep inference for *N-Hop* or *Intermediate Entity* tasks.

Figures 4–5 reveal a trade-off between capability and stability: while standard models converge immediately, reasoning-oriented models exhibit significant fragility. This instability escalates with unstructured context and task complexity—`OpenAI o4 Mini` surges from 1.53 to 5.00 attempts in noisy settings—identifying *N-Hop* reasoning as the primary bottleneck (Table 4). However, symbolic markup acts as a critical stabilizer, reducing retry rates for models like `DeepSeek-R1`, whereas outliers like `Kimi K2` prove consistently inefficient.

6 Discussion

This work highlights that automated mechanistic reasoning remains a challenging task where neuro-symbolic integration is critical. Our results demonstrate that performance peaks with symbolic markup rather than larger contexts, as unstructured text primarily acts as noise. Furthermore, we identify a fundamental trade-off between capability and stability: while reasoning-oriented models outperform baselines, they exhibit significant fragility, frequently triggering "failure loops" on complex queries. Symbolic markup resolves this tension as a "dual stabilizer," enhancing accuracy while drastically reducing convergence costs, suggesting that robust reasoning requires structured evidence rather than simply scaling model size.

7 Conclusion

We presented a scalable methodology for constructing large-scale biomedical mechanistic reasoning datasets, generating over 5 million semantically consistent 2-hop paths alongside a human-validated gold standard. Our benchmarking reveals that current LLMs struggle with multi-hop inference in broad contexts, exhibiting a stark trade-off between reasoning capability and operational stability. However, we demonstrate that neuro-symbolic scaffolding significantly mitigates this fragility, acting as a critical stabilizer that enhances both accuracy and convergence efficiency. By providing a reliable testbed for mechanistic logic, this work enables new directions for training robust neuro-symbolic models and advancing automated discovery in biomedical research.

Limitations

While our methodology enables the scalable construction of mechanistic datasets, it entails several limitations inherent to automated extraction and heuristic filtering.

Source and Extraction Constraints. Our pipeline relies on the REACH system, which is limited by data availability; subscription-based articles often lack full-text access, restricting extraction to abstracts and metadata. Furthermore, automated parsing introduces upstream errors, including incorrect sentence segmentation (e.g., periods within abbreviations breaking syntax trees) and the omission of specific document sections like methodology.

Granularity and Evaluation Ambiguity. A critical challenge in benchmarking reasoning is the definition of a "hop." Our gold standard assumes the REACH-extracted interactions are ground truth; however, biologically direct interactions may implicitly contain intermediate steps. Consequently, an LLM might correctly identify a latent intermediate entity in what we labeled as a 1-hop path, leading to a false negative in the evaluation. Finally, despite entity normalization, issues with duplicate IDs and nested entities (where one entity is a substring of another) may introduce residual noise in entity grounding.

References

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text. *arXiv preprint*. ArXiv:2403.18421 [cs].

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*. ArXiv:2211.12588 [cs].

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard A A Jonker, Sofia I R Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, and 1 others. 2024. The overview of the biored (biomedical relation extraction dataset) track at biocreative viii. *Database*, 2024:baae069.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1681–1701, Vienna, Austria. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 workshop companion volume for shared task*, pages 1–9.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. MedAdapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22294–22314, Miami, Florida, USA. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin

Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint*. ArXiv:2305.12031 [cs].

Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*. ArXiv:2203.11171 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Jiageng Wu, Xian Wu, and Jie Yang. 2024. Guiding clinical reasoning with large language models via knowledge seeds. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint*. ArXiv:2210.03629 [cs].

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2205.10625 [cs].

A Interaction Filtering


```
{
    "text": "It may stimulate the proliferation of
         skin fibroblasts through ERK signaling and
         may also promote osteoblast and mesenchymal
          cell differentiation [16].",
    "markup": "<span class=\"event
         Positive_activation\"> <span class=\"
         controller\"> It </span> may <span class=\"
         trigger\"> stimulate </span> the
         proliferation of <span class=\"controlled
         \"> skin fibroblasts </span> </span>
         through ERK signaling and may also promote
         osteoblast and mesenchymal cell
         differentiation [16].",
    "article": {
        "url": "https://www.ncbi.nlm.nih.gov/pmc/
             articles/PMC8033967/"
    },
    "controller": {
        "category": "uaz",
        "description": "It",
    },
    "controlled": {
        "category": "cl",
        "description": "skin fibroblasts"
    }
}
```


Figure 3: Example of a single biomedical interaction presented as an excerpt of the full JSON. The `text` field contains the original sentence from the article, while the `markup` field highlights the annotated event, including the `controller` ("It"), the `trigger` ("stimulate"), and the `controlled` entity ("skin fibroblasts"). The `article` field provides the source URL, and the `controller` and `controlled` fields include the entity category and description. Interactions with pronouns are removed, and only those where both entities are bioprocesses (*mesh*, *go*, *frailty*, *bioprocess*) or contain an **uppercase** letter or digit are retained for multihop path construction.

B Dataset Statistics

Table 1 shows the data volume statistics across construction stages, and Table 2 summarizes the number of resulting 1-hop and 2-hop paths in the dataset.

Table 1: Data volume statistics across construction stages. The dataset construction pipeline proceeds from document filtering to interaction extraction and path generation.

Stage	Count
Document Processing	
Original Source Documents	4.19M
Filtered Documents	350K
Interaction Extraction	
Raw Extractions Interactions	142,916
Validated 1-Hop Relations	107,263
Reasoning Path Construction	
Constructed 2-Hop Paths	$\sim$5M

Table 2: Distribution of reasoning paths across data splits.

Path	Split	Count	%
1-hop	Train	60,658	60
	Val	20,219	20
	Test	20,220	20
2-hop	Train	4,063,550	81.55
	Val	450,251	9.04
	Test	469,001	9.41
	Gold	137	–

C Evaluation and Experimental Setup

C.1 Gold Standard Annotation

To establish a high-quality benchmark for biomedical mechanistic reasoning, we randomly sampled 320 2-hop paths from the 2-hop test partition, ensuring that every selected instance contained unique entity pairs and interactions. We employed a double-blind annotation process involving a pool of four expert annotators, where each path was independently reviewed by two annotators. The process yielded substantial inter-annotator agreement, achieving a score of 0.72 on Cohen's κ. To prioritize precision, we filtered the dataset for strict unanimity, retaining only the instances where both annotators marked the path as correct. This rigorous selection process resulted in a final gold standard of 137 validated 2-hop paths for evaluation.

C.2 Evaluation Protocol

We benchmarked the reasoning capabilities of 11 Large Language Models (LLMs), comprising 7 reasoning-oriented architectures and 4 non-reasoning baselines. The evaluation probed four distinct aspects of mechanistic understanding using the following query types:

- **Relation Existence:** Verifying if a valid biological interaction exists between the two entities.
- **N-Hop Count:** Determining the path length (number of hops) between entities, assuming a relation exists.
- **Intermediate Entity:** Identifying the bridging entity within a 2-hop path.
- **Polarity:** Classifying the nature of the interaction (e.g., positive/negative).

We assessed model performance across five levels of contextual granularity to analyze the impact of noise and structure:

1. Target sentence with symbolic markup.
2. Target sentence without symbolic markup.
3. Sentence window (left context + target + right context).
4. Target paragraph.
5. Paragraph window (left context + target + right context).

For fair comparison, the maximum output length was fixed at 512 tokens for all models. To differentiate between reasoning failure and formatting failure, we allowed up to 10 retry attempts per query for the model to generate a valid, format-compliant response.

D Analysis of Performance, Inference Stability and Convergence

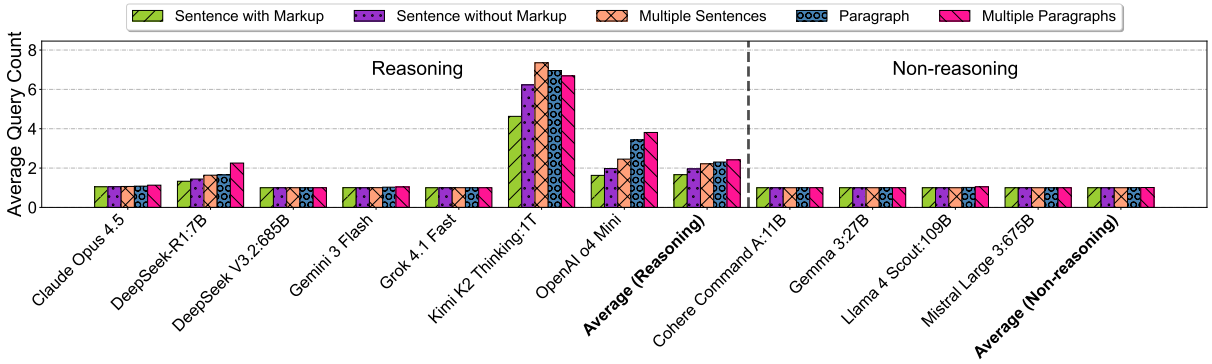

Figure 4: Impact of context complexity on inference stability. Non-reasoning models (right) demonstrate decisive, single-attempt convergence regardless of input size, whereas reasoning-oriented models (left) exhibit significant fragility. Specifically, for models like `OpenAI o4 Mini`, `DeepSeek-R1`, and `Kimi K2 Thinking`, unstructured contexts (e.g., *Multiple Paragraphs*) act as noise that triggers "inference thrashing", resulting in a monotonic increase in retry rates, whereas symbolic evidence (*Sentence with Markup*) consistently minimizes the computational cost of convergence.

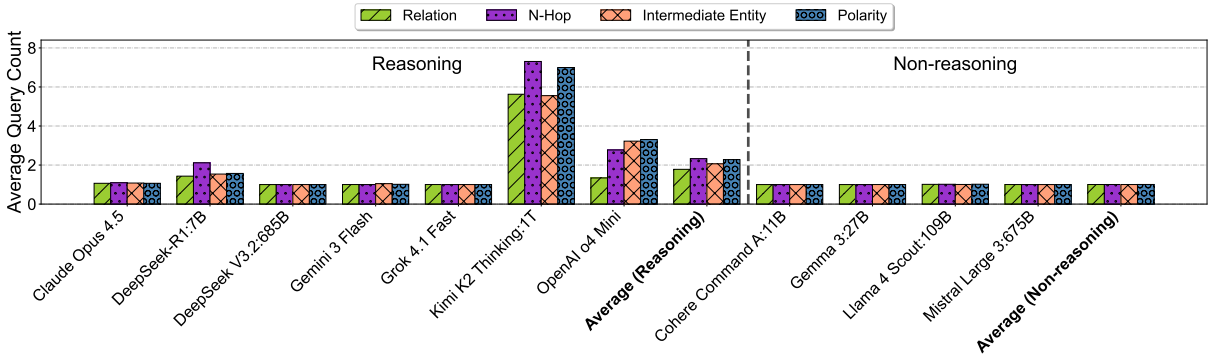

Figure 5: Impact of query complexity on inference stability. A comparison of average retry counts reveals that while non-reasoning models remain operationally static, reasoning models exhibit task-dependent instability. *N-Hop* reasoning consistently demands the most retries for models like `OpenAI o4 Mini`, `DeepSeek-R1`, and `Kimi K2 Thinking`, confirming that multi-step logic is the primary driver of failure loops in few-shot settings. Meanwhile, `Kimi K2 Thinking` proves consistently inefficient regardless of task type, approaching the retry limit across the board.

Table 3: Percentage of correct predictions across models, query types (*Relation*, *N-Hop*, *Intermediate Entity*, *Polarity*), and input granularities ranging from *Sentence with Markup* to *Multiple Paragraphs*. Models are ordered by reasoning capability, with the top models being reasoning-oriented and the bottom four being non-reasoning baselines. Accuracy generally decreases as structural cues are removed and context length increases, with reasoning models showing greater robustness—particularly for *N-Hop* queries.

Model	Query Type	Sentence with Markup	Sentence without Markup	Multiple Sentences	Paragraph	Multiple Paragraphs
Claude Opus 4.5	Relation	50.4	23.4	18.2	20.4	24.8
	N-Hop	**98.5**	70.1	54.7	34.3	35.0
	Intermediate Entity	**95.6**	83.9	77.4	53.3	47.4
	Polarity	89.1	77.4	74.5	66.4	67.9
DeepSeek-R1:7B	Relation	85.4	66.4	59.1	36.5	44.5
	N-Hop	75.2	54.0	48.2	16.1	17.5
	Intermediate Entity	73.0	54.0	48.2	24.8	11.7
	Polarity	73.7	75.2	67.2	57.7	54.0
DeepSeek V3.2:685B	Relation	67.9	34.3	28.5	29.2	27.0
	N-Hop	**91.2**	56.9	52.6	38.0	43.1
	Intermediate Entity	**93.4**	69.3	62.0	43.1	40.1
	Polarity	85.4	75.9	76.6	62.0	69.3
Gemini 3 Flash	Relation	**93.4**	65.7	60.6	49.6	57.7
	N-Hop	**96.4**	68.6	66.4	39.4	37.2
	Intermediate Entity	**98.5**	**87.6**	**81.8**	59.9	54.7
	Polarity	85.4	82.5	75.9	**68.6**	67.9
Grok 4.1 Fast	Relation	81.8	62.8	56.9	32.1	30.7
	N-Hop	**92.0**	69.3	66.4	40.1	40.1
	Intermediate Entity	**94.2**	78.8	75.2	56.9	57.7
	Polarity	85.4	73.7	75.2	67.2	65.7
Kimi K2 Thinking:1T	Relation	69.3	48.2	32.1	5.1	0.7
	N-Hop	65.0	36.5	18.2	2.2	0.7
	Intermediate Entity	85.4	67.9	54.0	14.6	6.6
	Polarity	59.9	39.4	35.8	10.9	3.6
OpenAI o4 Mini	Relation	61.3	48.9	35.8	15.3	15.3
	N-Hop	89.1	75.2	67.2	38.0	25.5
	Intermediate Entity	82.5	71.5	66.4	44.5	41.6
	Polarity	85.4	73.7	70.1	53.3	46.7
Cohere Command A:11B	Relation	56.2	48.2	38.7	36.5	35.8
	N-Hop	67.2	44.5	24.1	22.6	18.2
	Intermediate Entity	89.8	71.5	65.7	40.1	42.3
	Polarity	75.9	74.5	68.6	60.6	65.0
Gemma 3:27B	Relation	**90.5**	65.0	58.4	66.4	68.6
	N-Hop	86.1	59.9	37.2	34.3	24.8
	Intermediate Entity	**94.2**	72.3	62.0	40.1	33.6
	Polarity	73.0	72.3	74.5	66.4	**70.1**
Llama 4 Scout:109B	Relation	78.8	63.5	55.5	42.3	43.8
	N-Hop	67.2	46.7	26.3	16.1	13.1
	Intermediate Entity	**92.0**	75.2	67.2	42.3	39.4
	Polarity	76.6	75.9	73.0	65.7	69.3
Mistral Large 3:675B	Relation	73.0	45.3	48.2	47.4	44.5
	N-Hop	**91.2**	57.7	42.3	46.0	44.5
	Intermediate Entity	**92.0**	70.8	64.2	48.2	40.9
	Polarity	79.6	73.7	71.5	63.5	65.0

Table 4: Operational stability vs. reasoning depth. Average attempts required for convergence reveal a stark contrast: standard models (e.g., `Mistral Large 3`, `Gemma 3`) exhibit decisive, single-attempt behavior, whereas reasoning-oriented models (e.g., `Kimi K2 Thinking`, `OpenAI o4 Mini`) show increased instability and higher retry rates as context noise and task difficulty increase.

Model	Query Type	Sentence with Markup	Sentence without Markup	Multiple Sentences	Paragraph	Multiple Paragraphs
Claude Opus 4.5	Relation	1.07	1.00	1.07	1.07	1.13
	N-Hop	1.07	1.07	1.07	1.13	1.13
	Intermediate Entity	1.07	1.07	1.07	1.07	1.13
	Likely Polarity	1.00	1.07	1.07	1.07	1.13
DeepSeek-R1:7B	Relation	1.08	1.14	1.31	1.48	2.15
	N-Hop	1.69	1.93	2.48	1.95	2.53
	Intermediate Entity	1.17	1.43	1.26	1.70	2.13
	Likely Polarity	1.39	1.28	1.50	1.47	2.21
DeepSeek V3.2:685B	Relation	1.00	1.00	1.00	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.00	1.00	1.00
	Likely Polarity	1.00	1.00	1.00	1.00	1.00
Gemini 3 Flash	Relation	1.00	1.00	1.01	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.01	1.12	1.11
	Likely Polarity	1.01	1.00	1.00	1.00	1.05
Grok 4.1 Fast	Relation	1.00	1.00	1.01	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.00	1.00	1.00
	Likely Polarity	1.00	1.00	1.00	1.00	1.00
Kimi K2 Thinking:1T	Relation	4.41	5.53	6.66	6.28	6.09
	N-Hop	**5.54**	**7.58**	**9.04**	8.12	7.49
	Intermediate Entity	3.05	4.73	6.16	7.53	7.45
	Likely Polarity	5.52	7.08	7.54	**8.28**	**7.78**
OpenAI o4 Mini	Relation	1.30	1.15	1.34	1.39	1.52
	N-Hop	1.80	2.34	3.11	3.09	3.52
	Intermediate Entity	1.53	2.15	2.82	4.62	5.00
	Likely Polarity	1.88	2.26	2.55	4.63	5.20
Cohere Command A:11B	Relation	1.00	1.00	1.00	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.00	1.00	1.00
	Likely Polarity	1.00	1.00	1.00	1.00	1.01
Gemma 3:27B	Relation	1.00	1.00	1.00	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.00	1.00	1.01
	Likely Polarity	1.00	1.00	1.00	1.01	1.00
Llama 4 Scout:109B	Relation	1.00	1.00	1.00	1.01	1.05
	N-Hop	1.00	1.00	1.00	1.02	1.06
	Intermediate Entity	1.00	1.00	1.00	1.01	1.02
	Likely Polarity	1.00	1.00	1.00	1.03	1.06
Mistral Large 3:675B	Relation	1.00	1.01	1.00	1.00	1.00
	N-Hop	1.00	1.00	1.00	1.00	1.00
	Intermediate Entity	1.00	1.00	1.00	1.00	1.00
	Likely Polarity	1.00	1.00	1.00	1.00	1.00