BioHopR: A Benchmark for Multi-Hop, Multi-Answer Reasoning in Biomedical Domain

Anonymous ACL submission

Abstract

Biomedical reasoning often requires traversing interconnected relationships across entities such as drugs, diseases, and proteins. Despite the increasing prominence of large language models (LLMs), existing benchmarks lack the ability to evaluate multi-hop reasoning in the biomedical domain, particularly for queries involving *one-to-many* and *many-to-many* relationships. This gap leaves the critical challenges of biomedical multi-hop reasoning underexplored. To address this, we introduce **BioHopR**, a novel benchmark designed to evaluate multi-hop, multi-answer reasoning in structured biomedical knowledge graphs. Built from the comprehensive PrimeKG, BioHopR includes 1-hop and 2-hop reasoning tasks that reflect real-world biomedical complexities.

Evaluations of state-of-the-art models reveal that O3-mini, a proprietary reasoning-focused model, achieves **37.93%** accuracy on 1-hop tasks and **14.57%** on 2-hop tasks, outperforming proprietary models such as GPT4O and open-source biomedical models including HuatuoGPT-o1-70B and Llama-3.3-70B. However, all models exhibit significant declines in multi-hop performance, underscoring the challenges of resolving implicit reasoning steps in the biomedical domain. By addressing the lack of benchmarks for multi-hop reasoning in biomedical domain, BioHopR sets a new standard for evaluating reasoning capabilities and highlights critical gaps between proprietary and open-source models while paving the way for future advancements in biomedical LLMs.

1 Introduction

Recent advances in large language models (LLMs) and Question Answering (QA) systems have shifted the focus from simple factoid retrieval tasks to more sophisticated reasoning capabilities (Huang and Chang, 2022; Plaat et al., 2024; OpenAI, 2025). Among these, **multi-hop reasoning** has emerged as a critical area of research, where answering a question requires traversing multiple interconnected reasoning steps (Misra et al., 2023; Yang et al., 2024; Schnitzler et al., 2024). For example, to answer *"Who is the wife of the president of the United States?"*, a LLM must first identify the president (step 1) and then determine their spouse (step 2). This type of reasoning, referred to as multi-hop reasoning, is especially vital in domains where information is highly interconnected, such as the biomedical field.

In the biomedical domain, knowledge is often structured in ontologies and knowledge graphs (KGs), where entities like drugs, diseases, proteins, and phenotypes are represented as nodes, and their relationships as edges (Himmelstein et al., 2017; Sung et al., 2021; Chandak et al., 2023). Biomedical queries frequently demand multi-step reasoning over these graphs (Sung et al., 2021; Su et al., 2024; Matsumoto et al., 2025). For instance, identifying diseases associated with a drug might require a single-hop relation, while determining proteins targeted by that drug through its associated disease involves two reasoning steps. Additionally, biomedical reasoning often entails *one-to-many* or *many-to-many* relationships, where a single query may yield multiple valid answers (e.g., a drug targeting multiple proteins) (Liang et al., 2019). This complexity highlights the need for specialized benchmarks that rigorously evaluate models' ability to reason across multiple steps while generating comprehensive, multi-answer responses.

Existing benchmarks for multi-hop reasoning, such as Hetionet (Himmelstein et al., 2017) and other biomedical QA datasets (Rao et al., 2022), have laid the groundwork for evaluating multi-hop capabilities in the biomedical domain. However, these benchmarks primarily focus on single-hop tasks or utilize pre-defined templates that fail to fully capture the intricacies of multi-step reasoning. Similarly, general-domain benchmarks like TWOHOPFACT (Yang et al., 2024) test mod-

els' latent multi-hop reasoning ability but lack the domain-specific challenges of biomedical reasoning, such as reasoning over structured relationships and handling multi-answer outputs. As a result, the unique challenges of biomedical multi-hop reasoning remain underexplored.

To address these limitations, we introduce **BioHopR**, a new benchmark specifically designed to test the multi-hop reasoning capabilities of LLMs in the biomedical domain. Unlike general-domain benchmarks that rely on reasoning across disconnected documents, BioHopR focuses on reasoning within a single, structured biomedical knowledge graph. Our benchmark systematically constructs 1-hop (e.g., Drug–Disease) and 2-hop (e.g., Drug–Disease–Protein) question-answer pairs from the PrimeKG knowledge graph (Chandak et al., 2023). Questions are designed to evaluate models' abilities to reason step-by-step, explicitly requiring the inference of intermediate entities, and generate multi-answer responses reflective of real-world biomedical complexity.

Contributions. Our main contributions are as follows:

- **A New Benchmark for Multi-Hop Reasoning:** We propose **BioHopR**, the first publicly available benchmark explicitly designed to evaluate multi-hop, multi-answer reasoning within structured biomedical knowledge graphs. We will release the dataset and the code for evaluation.
- **Evaluation and Analysis of LLMs in Biomedical Multi-hop Reasoning:** We evaluate state-of-the-art LLMs on our benchmark, highlighting their strengths and, more importantly, limitations in handling biomedical multi-hop reasoning tasks.

By introducing BioHopR, we aim to fill a critical gap in multi-hop QA research and advance the development of LLMs capable of robust and interpretable reasoning in structured, high-stakes domains like biomedical research and healthcare.

2 Related Works

Biomedical Question Answering. Research in medical LLMs has been facilitated by the development of question-answering (QA) datasets that benchmark models' understanding of medical domain knowledge (Hendrycks et al., 2020; Jin et al., 2021; Pal et al., 2022). These datasets typically consist of multiple-choice questions (MCQs) focused on single-hop reasoning tasks, providing a straightforward way to evaluate LLMs' ability to comprehend and respond to diverse medical inquiries. While these benchmarks have driven significant progress, they primarily measure classification accuracy, which is insufficient for capturing the nuanced reasoning required for medical expertise.

Medical QA often involves interconnected concepts where reasoning over multiple steps is crucial. However, current benchmarks rarely go beyond single-hop tasks and do not evaluate models' ability to provide explanations for their answers or justify their reasoning process. Recently, MedExQA introduced an evaluation framework with detailed explanations for assessing the reasoning capabilities of LLMs (Kim et al., 2024). While this is a step forward, it remains constrained to single-hop reasoning and does not address the need for multi-hop reasoning or the generation of multiple valid answers—a common requirement in biomedical inquiries.

Knowledge Graph Question Answering. Knowledge Graph Question Answering (KGQA) systems leverage structured knowledge graphs to answer questions that require reasoning over graph-based relationships. In the biomedical domain, Hetionet (Himmelstein et al., 2017) introduced a knowledge graph containing entities like genes, drugs, and diseases, enabling structured reasoning. Extensions of Hetionet have been used for multi-hop QA tasks (Rao et al., 2022), but these datasets often rely on fixed templates and predefined reasoning paths, limiting their ability to evaluate the nuanced multi-hop reasoning required in real-world biomedical applications. This work explored techniques such as knowledge graph embeddings and graph neural networks (Kipf and Welling, 2016; Hamilton et al., 2018), and transformer-based models like BioBERT (Lee et al., 2020) to extract and utilize graph-based knowledge. However, this dataset tests the model's performance in a classification task to a single answer. Also, this dataset is not publicly available, limiting its role in facilitating biomedical large language model research.

In domains like biomedical science, many questions inherently involve multiple correct answers. For instance, identifying all drugs that treat a specific disease or all proteins associated with a dis-

Dataset	Domain	Reasoning	Answer-Level
MedQA (Jin et al., 2021)	Biomedical	No	Single Answer
Hetionet QA (Himmelstein et al., 2017)	Biomedical	Graph-based Reasoning (MH)	Single Answer
MedExQA (Kim et al., 2024)	Biomedical	Explanation Generation	Single Answer
TWOHOPFACT (Yang et al., 2024)	General	Implicit Reasoning (MH)	Single Answer
BioHopR (Ours)	Biomedical	Implicit Reasoning (MH)	Multi Answers

Table 1: Comparison of BioHopR with existing datasets. Key differentiators include domain focus, reasoning type (MH is tagged for multi-hop reasoning supported dataset), and answer-level, such as multi-answer capability.

ease phenotype requires models to retrieve comprehensive sets of answers rather than a single response.

Latent Multi-Hop Reasoning in Large Language Models. Recent work has explored the latent reasoning capabilities of LLMs, focusing on whether models can implicitly infer intermediate entities and use them for multi-step reasoning. The TWOHOPFACT dataset (Yang et al., 2024) evaluates this capability by testing whether LLMs can identify "bridge entities" in two-hop reasoning tasks. While TWOHOPFACT demonstrates that LLMs can perform latent multi-hop reasoning in general domains, it does not address the unique challenges of biomedical reasoning. Biomedical queries often require explicit reasoning over structured data and demand comprehensive answers involving one-to-many or many-to-many relationships.

These gaps highlight the need for a benchmark like BioHopR, which explicitly evaluates models' ability to perform step-by-step reasoning and generate multi-answer outputs in the biomedical domain.

Multi-Answer Reasoning. Existing QA benchmarks, both in general and biomedical domains, typically assume a one-to-one mapping between questions and answers, which oversimplifies the complexity of real-world reasoning tasks. This assumption is especially problematic in the biomedical domain, where relationships between entities are often one-to-many or many-to-many.

BioHopR BioHopR addresses this limitation by introducing questions that require multi-answer reasoning, ensuring that the benchmark captures the intricate relational structures and knowledge dependencies present in biomedical science. The differences between our dataset and relevant datasets are summarized in Table 1.

3 BioHopR: Multi-hop Reasoning in Biomedicine

BioHopR is a benchmark specifically designed to evaluate the ability of large language models (LLMs) in performing multi-hop reasoning and generating multi-answer outputs in the biomedical domain. Compared to other knowledge graphs such as Hetionet, PrimeKG provides a broader coverage of biomedical entities, richer relational structures, and up-to-date knowledge in the field (Chandak et al., 2023). This allows for the generation of diverse, clinically relevant up-to-date multi-hop queries. By systematically constructing questions over *PrimeKG*, the dataset is constrained to follow a **one-many-many** relationship structure. This restriction ensures that queries reflect real-world biomedical scenarios where entities like drugs, diseases, and proteins exhibit hierarchical and complex interconnected relationships.

3.1 Multi-hop, Multi-answer Knowledge Formalization

Nodes and Relations. In our dataset, the entities in PrimeKG are represented as nodes, and their relationships are directed edges. For any query, the node from which reasoning starts is defined as the **query node**, and the node(s) forming the final answers are the **target nodes**. In the case of 2-hop reasoning, the intermediate node connecting the query and target is defined as the **bridge node**. We restrict node types to the following: Drug, Proteins, Disease, Phenotype.

Relationship Structure. The dataset is restricted to follow a **one-many-many** relationship structure.

In 1-hop questions, a direct relationship connects the **query node** to the **target nodes**. For example:

$$\text{Query (Drug)} \xrightarrow{\text{treats}} \text{Target (Diseases)}. \quad (1)$$

This setup reflects a single reasoning step where a query node is linked to multiple target nodes.

In 2-hop questions, the **query node** connects to the **target nodes** via an intermediate **bridge node**, forming a two-step reasoning chain. For example:

$$\text{Query (Phenotype)} \xrightarrow{\text{side\_effects\_of}} \text{Bridge (Drug)} \xrightarrow{\text{treats}} \text{Target (Diseases)}. \quad (2)$$

Here, the bridge node (e.g., drug), used to query for 1-hop questions, serves as the intermediate entity linking the query and target.

Answer Definition. The **target nodes** are the final answers to the query. For 1-hop reasoning, this corresponds to all nodes directly connected to the query node. For 2-hop reasoning, the answers are all nodes reachable via the bridge node, requiring models to infer both the intermediate (bridge) and final (target) nodes.

3.2 Dataset Construction Pipeline

The dataset is constructed using the following systematic process:

1. **Entity Sampling:** Nodes representing drugs, diseases, proteins, and phenotype entities are extracted from PrimeKG.
2. **2-Hop Path Definition:** For 2-hop questions, valid paths are constructed by combining two connected edges, ensuring the **query-bridge-target** structure follows the one-many-many relationship:

$$\text{Query} \xrightarrow{\text{Relation}_1} \text{Bridge} \xrightarrow{\text{Relation}_2} \text{Target}. \quad (3)$$

3. **1-Hop Relationship Extraction:** For 1-hop questions, all relationships connecting query nodes (e.g., drugs) to their target nodes (e.g., diseases) are extracted. To maintain consistency with 2-hop questions, 1-hop relationships without a corresponding 2-hop path are excluded.
4. **Answer Extraction:** For each question, all reachable target nodes are extracted as answers. This ensures that the multi-answer nature of the dataset is preserved.

3.3 1-Hop and 2-Hop Questions

1-Hop Questions. For a 1-hop question, the model is required to directly link the query node to the target node, such as:

"Name a disease that is treated by Drug Dr?" (4)

with the answer set defined as:

$$A = \{D_1, D_2, \ldots, D_n\}, \quad (5)$$

where D_i represents disease linked to the query drug Dr.

2-Hop Questions. For a 2-hop question, the model must infer both the bridge node and the target node. An example query is:

"Name a diseases that is treated by a drug that has a side effect$S$?" (6)

The model needs to traverse the graph through an intermediate **bridge node** (drug) before reaching the final **target node** (disease):

$$A = \{D_1, D_2, \ldots, D_n\}, \quad (7)$$

where D_i represents disease linked to the phenotype that is a side effect of drug Dr.

3.4 Dataset Statistics

Relation (Query:Target)	Count
Protein:Disease	731
Protein:Drug	589
Disease:Drug	297
Drug:Phenotype	248
Drug:Disease	234
Drug:Protein	165
Disease:Protein	113
Disease:Phenotype	79
Phenotype:Drug	33
Phenotype:Disease	5

Table 2: Distribution of 1-hop relations in BioHopR.

The **BioHopR** dataset consists of **2,494** unique 1-hop questions and **7,633** unique 2-hop questions, resulting in a total of **279,738** answers. On average, each question is associated with **36.65** answers, reflecting the dataset's complexity and the many-to-many relationships inherent in biomedical knowledge. The dataset includes 10 distinct 1-hop relation types and 12 2-hop relation types, and the breakdown of the number of questions for each relation type is summarized in Tables 2 and 3.

The restriction to one-many-many relationships ensures that the dataset mirrors real-world biomedical reasoning scenarios, where single entities often relate to multiple downstream entities. This design makes the dataset uniquely suited for evaluating large language models (LLMs) on tasks requiring multi-step reasoning and comprehensive answer generation.

Relation (Query:Bridge:Target)	Count
Drug:Protein:Disease	3029
Disease:Drug:Phenotype	949
Disease:Protein:Drug	899
Protein:Disease:Drug	577
Phenotype:Disease:Drug	546
Protein:Drug:Disease	462
Disease:Drug:Protein	381
Drug:Disease:Protein	321
Phenotype:Drug:Disease	215
Drug:Disease:Phenotype	213
Disease:Phenotype:Drug	36
Drug:Phenotype:Disease	5

Table 3: Distribution of 2-hop relations in BioHopR.

3.5 Qualitative Analysis

To better understand the models' reasoning capabilities, we conducted a qualitative analysis on the questions about Type II Diabetes, as it is one of the widely studied diseases (Skyler et al., 2017).

3.6 Reasoning Benchmark

BioHopR presents significant reasoning challenges:

- Models must implicitly identify intermediate bridge nodes in 2-hop questions while ensuring the correctness of the final answers.
- The many-to-many nature of biomedical relationships requires models to handle diverse answer sets while preserving reasoning consistency.

4 Experiments

We evaluate a range of LLMs on the BioHopR benchmark to assess their ability to reason over one-many-many relationships. The evaluation focuses on both single-answer and multi-answer reasoning for 1-hop and 2-hop questions, highlighting the challenges posed by multi-step reasoning and comprehensive answer generation.

4.1 Experimental Setup

Models Evaluated. We consider a diverse set of LLMs, categorized into general-purpose proprietary, reasoning proprietary, medical-specific, and open-source models, as detailed in Table 4. General-purpose models include GPT4O and smaller variants such as GPT4O-mini (Hurst et al., 2024). We also added O3-mini as it was most recent cost-effective reasoning proprietary model (OpenAI, 2025). We also evaluate open-source Llama models (Llama3.1 and Llama3.3) with varying parameter scales (8B and 70B) (Dubey et al., 2024). We selected medical-specific models that are based on the baseline Llama3.1 architectures: UltraMedical-8B, HuatuoGPT-o1-8B, and HuatuoGPT-o1-70B (Zhang et al., 2024; Chen et al., 2024). HuatuoGPT-o1 models are trained for medical complex reasoning for medical problems.

Model Name	Domain
GPT4O	General
GPT4O-mini	General
O3-mini	Reasoning
Llama3.1 8B	General
Llama3.1 70B	General
Llama3.3 70B	General
UltraMedical-8B	Medical
HuatuoGPT-o1-8B	Medical Reasoning
HuatuoGPT-o1-70B	Medical Reasoning

Table 4: Models evaluated in the experiments.

4.2 Evaluation

The proprietary GPT models (GPT4O, GPT4O-mini, and O3-mini) were accessed using OpenAI's API[1]. For open-source models, we used four A100 GPUs with 80GB memory per GPU for 70B parameter models and one A6000 GPU for 8B parameter models. The evaluation was conducted in a zero-shot setting, with a batch size of 1 and a temperature set to 0, except O3-mini model which does not support temperature parameter, to ensure deterministic responses. The evaluation code for open-source models were implemented using the HuggingFace Transformers library (Wolf, 2019).

4.3 Evaluation Metrics

Embedding-Based Accuracy. Accuracy (**ACC**) is computed using the cosine similarity between the predicted response and the ground truth answer list, leveraging BioLORD-2023-C embeddings (Remy et al., 2023). Let p denote the embedding of the predicted response and $\{a_1, a_2, \ldots, a_n\}$ denote the embeddings of the ground truth answers. The cosine similarity for a prediction p and an answer a_i is defined as:

$$\cos(p, a_i) = \frac{p \cdot a_i}{\|p\| \|a_i\|}. \quad (8)$$

[1] https://platform.openai.com/docs/models

Model	ACC_HOP1 (%)	ACC_HOP2 (%)	BOTH_COR (%)	BOTH_WR (%)
Llama-3.1-8B	0.12	0.05	0.00	99.76
HuatuoGPT-o1-70B	0.16	0.00	0.00	99.93
HuatuoGPT-o1-8B	0.20	0.04	0.00	99.54
UltraMedical-8B	13.75	5.21	2.28	82.33
Llama-3.3-70B	25.58	9.58	4.94	68.33
Llama-3.1-70B	26.38	9.47	4.93	65.64
GPT4O-mini	28.11	14.57	6.54	64.69
GPT4O	32.88	14.57	7.86	57.96
O3-mini	37.93	14.57	8.93	52.14

Table 5: Performance metrics (in percentages) for various models. ACC_HOP1 and ACC_HOP2 represent the accuracy on 1-hop and 2-hop tasks, respectively. BOTH_COR indicates cases where both hops are correct, and BOTH_WR indicates cases where both hops are incorrect.

If the maximum cosine similarity across all ground truth answers satisfies:

$$\max_{i \in \{1,\ldots,n\}} \cos(p, a_i) > \tau, \tag{9}$$

then the prediction is considered correct. We use $\tau = 0.9$ for BioLORD-2023-C embeddings after a grid search of threshold values from 0.5 to 0.9, which led an optimal setting with 0.9. This threshold prioritizes precision, ensuring that only highly confident predictions are accepted as correct. By setting a high threshold, we align with the strict requirements of biomedical applications, minimizing false positives while maintaining robust handling of biomedical definition-level similarity and ambiguous synonyms.

5 Results and Discussion

5.1 Proprietary Models Demonstrate Robust Multi-Hop Reasoning

Proprietary models (GPT4O, GPT4O-mini, and O3-mini) demonstrate consistently strong performance across all metrics. For 1-hop tasks (**ACC_HOP1**), O3-mini achieves the highest accuracy (37.93%), followed by GPT4O (32.88%) and GPT4O-mini (28.11%). Interestingly, all proprietary models achieve identical performance on 2-hop tasks (**ACC_HOP2: 14.57%**), suggesting a possible shared capabilities for implicit reasoning or complex reasoning.

These results reflect the impact of the reasoning step before answering. O3-mini's higher **ACC_HOP1** indicates the reasoning capability of the model allowed it to reason well on single-step queries.

5.2 Open-Source Biomedical Models Face Significant Challenges

Open-source biomedical models struggle to match the performance of proprietary models, particularly on multi-hop tasks. HuatuoGPT-o1 models perform the worst, achieving near-zero accuracy for both 1-hop (**ACC_HOP1: 0.20%** for HuatuoGPT-o1-8B) and 2-hop (**ACC_HOP2: 0.00%** for HuatuoGPT-o1-70B). In contrast, UltraMedical-8B performs better (**ACC_HOP1: 13.75%**, **ACC_HOP2: 5.21%**).

These results suggest that although HuatuoGPT-1 was trained for medical complex reasoning, it's generalizability is far less than UltraMedical. The reasoning demands of BioHopR is far different from medical license examination based QA datasets such as MedQA, which HuatuoGPT-o1 used for training. Still UltraMedical-8B's performance, when compared to a larger general domain open-source models such as Llama3.1-70B and Llama3.3-70B, is far behind, suggesting persistent challenges in resolving bridge nodes for multi-hop queries.

Error Patterns. The **BOTH_WR** metric reveals systemic challenges in multi-hop reasoning for all models. Open-source models like HuatuoGPT-o1-70B exhibit the highest **BOTH_WR** rates (>99%), reflecting widespread failure in both reasoning hops. Proprietary models demonstrate significantly lower failure rates, with O3-mini achieving the best performance (**BOTH_WR: 52.14%**). However, even the best-performing models show substantial error rates in both hops, indicating that multi-step inference remains a bottleneck.

5.3 Multi-Hop Reasoning Remains a Bottleneck

Across all models, performance declines sharply from 1-hop to 2-hop tasks. For example, GPT4O's accuracy drops from **ACC_HOP1: 32.88%** to **ACC_HOP2: 14.57%**, while open-source models like Llama-3.1-8B exhibit near-complete failure (**ACC_HOP2: 0.05%**).

This decline highlights the inherent complexity of multi-hop reasoning. Resolving 2-hop queries requires implicit inference of intermediate entities (e.g., bridge nodes) and alignment of reasoning chains across multiple steps.

5.4 Qualitative Analysis - Case Studies

Our qualitative analysis on various diseases, including Type II Diabetes and Schizophrenia aligns well with the evaluation result in Table 5. We highlight the diabetes-related questions in Figure 1. Diabetes-related questions for drug Troglitazone, which has 202 side effects listed from PrimeKG, highlighted mixed performance among models. For instance, HuatuoGPT-o1-8B correctly predicted answers but diverged from the task constraints by elaborating on its reasoning instead of adhering to the prompt. Similarly, UltraMedical produced multiple answers when a single response was requested, with only some of the predictions being correct. In contrast, proprietary models such as GPT-4 reliably adhered to prompted task, consistently including relevant responses such as hepatotoxicity, even if these were not explicitly part of the predefined answer set. This behavior suggests that proprietary models may apply broader medical reasoning compared to open-source models. Proprietary models generally outperform open-source models in both task adherence and reasoning accuracy.

Questions		
Hop1: *"...a side effect of drug Troglitazone."* **Hop2**: *"...a side effect of a drug ... treat type 2 diabetes."*		
Model	**Hop1 Prediction**	**Hop2 Prediction**
HuatuoGPT-o1-70B	*"Alright, let's think about Troglitazone..."*	*"Alright, let's think about this..."*
HuatuoGPT-o1-8B	*"Hepatotoxicity"*	*"Hypoglycemia"*
UltraMedical-8B	*"Hepatotoxicity"*	*"Lactic acidosis, Hypoglycemia, Hyperkalemia"*
GPT4O	*"Hepatotoxicity"*	*"Weight gain"*
O3-mini	*"Hepatotoxicity"*	*"Weight gain"*

Figure 1: Qualitative analysis of model responses to diabetes-related questions. Red-colored text shows the wrong answer. Orange-colored text shows the answer that is not in the answer list, but is plausible. Blue-colored text shows the correct answer.

5.5 Ablation Study: Prompting Strategy

Prompting Setup. The dataset can also support multi-answer prompting, making two prompting strategies designed to evaluate different aspects of model reasoning:

- **Single-Answer Prompting:** The model is prompted to provide one correct answer (e.g., *"Name a gene associated with Disease X."*). This evaluates the model's ability to identify the most probable answer using implicit reasoning.
- **Multi-Answer Prompting:** The model is prompted to provide all correct answers (e.g., *"Name all genes associated with Disease X."*). This evaluates the model's ability to generate exhaustive, comprehensive outputs, which is inherently more challenging.

While both strategies are valuable for understanding model performance, **Multi-Answer Prompting** poses significant challenges. On average, each question in the dataset has 36.65 correct answers, making it computationally expensive and cognitively demanding for large language models to generate a complete answer set.

5.5.1 Evaluation Metric for Multi-Answer

F1 Score for Multi-Answer Prompting. For **Multi-Answer Prompting**, F1 score is computed using cosine similarity-based matching. Let $P = \{p_1, p_2, \ldots, p_m\}$ denote the embeddings of the predicted responses and $A = \{a_1, a_2, \ldots, a_n\}$ denote the embeddings of the ground truth answers. A predicted response p_j is considered a true positive if:

$$\max_{i \in \{1,\ldots,n\}} \cos(p_j, a_i) > \tau, \tag{10}$$

where we set $\tau = 0.9$ for high-confidence matches. For **Single-Answer Prompting**, we use the same evaluation metric, **ACC**.

5.5.2 Analysis and Results.

To analyze the feasibility of Multi-Answer Prompting, we conducted an ablation study using GPT4O and GPT4O-mini, the proprietary models. Table 6 presents the performance metrics for Single-

Relation Type	1-Hop		2-Hop	
	GPT4O (ACC / F1)	GPT4O-mini (ACC / F1)	GPT4O (ACC / F1)	GPT4O-mini (ACC / F1)
Same Query and Bridge				
`Disease:Drug:Phenotype`	47.47 / 35.35	43.77 / 29.21	25.08 / 8.20	25.08 / 7.21
`Disease:Drug:Protein`	47.47 / 35.35	43.77 / 29.21	3.67 / 0.22	3.67 / 1.27
`Drug:Disease:Phenotype`	55.13 / 14.41	50.85 / 16.70	22.07 / 9.75	22.07 / 10.19
`Drug:Disease:Protein`	55.13 / 14.41	50.85 / 16.70	4.67 / 0.48	4.67 / 0.65
Same Query and Target				
`Disease:Phenotype:Drug`	20.25 / 9.22	22.78 / 8.51	16.67 / 4.04	16.67 / 2.73
`Disease:Protein:Drug`	35.40 / 1.63	26.55 / 4.82	8.12 / 3.76	8.12 / 3.76
`Drug:Phenotype:Disease`	23.39 / 8.62	31.05 / 6.71	0.00 / 0.00	0.00 / 0.00
`Drug:Protein:Disease`	20.61 / 2.31	20.00 / 15.36	20.14 / 4.54	20.14 / 4.54
Others				
`Phenotype:Disease:Drug`	0.00 / 0.05	0.00 / 3.49	14.47 / 6.63	14.47 / 6.63
`Phenotype:Drug:Disease`	15.15 / 7.63	24.24 / 4.10	3.72 / 2.77	3.72 / 2.77
`Protein:Disease:Drug`	35.29 / 10.06	27.50 / 8.22	2.95 / 2.54	2.95 / 2.54
`Protein:Drug:Disease`	23.60 / 9.46	14.43 / 10.84	1.08 / 1.49	1.08 / 1.49
Overall	**32.88** / 7.29	28.11 / **9.41**	14.57 / 4.96	14.57 / **5.22**

Table 6: Comparison of Single-Answer prompting (**ACC**) and Multi-Answer prompting (**F1**) for GPT4O and GPT4O-mini across 1-hop and 2-hop relation types.

Answer and Multi-Answer prompting across 1-hop and 2-hop tasks.

Single-Answer Prompting Outperforms Multi-Answer Prompting: GPT4O achieves an average ACC of 32.88% in 1-hop tasks, significantly higher than its Multi-Answer F1 score of 7.29%. The gap is even more pronounced in 2-hop tasks, where GPT4O achieves 14.57% ACC compared to just 4.96% F1. Relations with abstract or less structured targets (e.g., `Disease:Drug:Protein`) exhibit particularly poor F1 scores under Multi-Answer prompting, with GPT4O achieving only 0.22% F1 in 2-hop tasks. These results highlight the difficulty of generating comprehensive answer sets, especially for complex reasoning paths.

Based on these findings, we restricted our evaluation of all other models to Single-Answer Prompting. This decision is motivated by higher robustness and computational overhead of multi-answer prompting. Also in many real-world scenarios, users typically seek the most probable or relevant answer, aligning more closely with Single-Answer prompting.

While Multi-Answer prompting offers valuable insights into a model's ability to generate exhaustive outputs, it remains a challenging evaluation paradigm. Future work could focus on improving model training and prompting strategies to better support comprehensive answer generation.

6 Conclusion

We introduced **BioHopR**, a benchmark for evaluating multi-hop, multi-answer reasoning in the biomedical domain. Built on the PrimeKG knowledge graph, BioHopR captures the complexity of real-world biomedical queries through *one-to-many* and *many-to-many* relationships, rigorously assessing reasoning over 1-hop and 2-hop tasks.

Evaluation results highlight that O3-mini, a proprietary model with a reasoning step, outperforms open-source models including biomedical models like HuatuoGPT-o1. Across all models, the performance drop from 1-hop to 2-hop tasks underscores the difficulty of aligning intermediate reasoning steps, especially in bridging entities.

By addressing the lack of benchmarks for multi-hop reasoning in biomedical domain, BioHopR sets a new standard for evaluating reasoning capabilities and provides a critical step toward more robust and interpretable LLMs for biomedical research and real-world applications. Future directions include expanding the dataset to other knowledge sources and domains, such as chemistry.

Limitation

While BioHopR provides a rigorous benchmark for evaluating multi-hop reasoning in the biomedical domain, several limitations exist. BioHopR is currently focused on 4 major entities only: Protein, Phenotype, Drug, Disease. Also, it relies exclusively on a single knowledge graph, PrimeKG, which, while comprehensive, may not fully capture the diversity of biomedical knowledge or its real-world dynamics. This lack of diversity could bias model evaluation toward the structure and content of 4 major node types and PrimeKG, potentially under-representing a model's ability to generalize to other knowledge and sources. While human evaluation was not the primary focus of this work, future efforts could include more extensive and diverse human evaluations to validate model-generated outputs.

Broader Impacts and Ethics Statement

Our work raises no major ethical concerns. All evaluations and experiments were conducted strictly for research purposes.

We will release BioHopR. License and copyright information, along with Terms of Use, will be made available upon release of the dataset and associated materials. While BioHopR facilitates advancements in biomedical reasoning tasks, it is not designed for use in real-world clinical applications. Consequently, models evaluated or trained on BioHopR should not be used for clinical decision-making without rigorous validation and regulatory approvals.

This restriction aims to mitigate potential risks associated with incorrect reasoning or hallucinated outputs, which could lead to harmful clinical outcomes. Additionally, while BioHopR supports research into biomedical reasoning, it is critical that researchers use the benchmark responsibly, with appropriate safeguards in place to ensure the ethical use of derived insights and outputs.

References

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Embedding logical queries on knowledge graphs. *Advances in neural information processing systems*, 31.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6:e26726.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiaomin Liang, Daifeng Li, Min Song, Andrew Madden, Ying Ding, and Yi Bu. 2019. Predicting biomedical relationships using the knowledge and graph embedding cascade model. *PLoS One*, 14(6):e0218264.

Nicholas Matsumoto, Hyunjun Choi, Jay Moran, Miguel E Hernandez, Mythreye Venkatesan, Xi Li, Jui-Hsuan Chang, Paul Wang, and Jason H Moore. 2025. Escargot: an ai agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics*, 41(2):btaf031.

Kanishka Misra, Cicero Nogueira dos Santos, and Siamak Shakeri. 2023. Triggering multi-hop reasoning for question answering in language models using soft prompts and random walks. *arXiv preprint arXiv:2306.04009*.

OpenAI. 2025. Openai o3-mini. https://openai.com/index/openai-o3-mini/. Accessed: 2025-02-01.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

Dattaraj J Rao, Shraddha S Mane, and Mukta A Paliwal. 2022. Biomedical multi-hop question answering using knowledge graph embeddings and language models. *arXiv preprint arXiv:2211.05351*.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. Biolord-2023: Semantic textual representations fusing llm and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*.

Jay S Skyler, George L Bakris, Ezio Bonifacio, Tamara Darsow, Robert H Eckel, Leif Groop, Per-Henrik Groop, Yehuda Handelsman, Richard A Insel, Chantal Mathieu, et al. 2017. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, 66(2):241–255.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. Knowledge graph based agent for complex, knowledge-intensive qa in medicine. *arXiv preprint arXiv:2410.04660*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Appendix

The Figure 2 illustrates the frequency distribution of target and bridge entities within the BioHopR dataset, highlighting key patterns. The left panel demonstrates the prevalence of proteins (e.g., CYP3A4), phenotypes (e.g., Nausea), drugs (e.g., Olanzapine), and diseases (e.g., Schizophrenia) as target nodes in multi-hop queries. Meanwhile, the right panel showcases the distribution of bridge entities, which frequently include proteins (e.g., CDK2), phenotypes (e.g., Neoplasm of the skin), drugs (e.g., Fostamatinib), and diseases. These patterns reflect the diversity and real-world complexity of biomedical entities, emphasizing the challenges of reasoning over structured knowledge graphs for multi-hop queries.

Figure 3 illustrates the results of a grid search for determining the optimal cosine similarity threshold for BioLORD-2023-C embeddings. The x-axis represents the threshold values, ranging from 0.1 to 0.9, while the y-axis shows the accuracy for "Both Correct" predictions. A sharp decline in accuracy is observed as the threshold increases, with accuracy plateauing beyond 0.8. The chosen threshold of 0.9 ensures high precision by accepting only highly confident predictions, aligning with the strict requirements of biomedical reasoning tasks.

A Detailed Qualitative Analysis

We further included other diseases: Vitamin A Deficiency, Lung Cancer, Alzheimer's Disease, Schizophrenia. We selected these medical conditions because they represent a range of domains within the biomedical field, which include nutritional deficiencies, metabolic disorders, chronic diseases and neurodegenerative conditions. This selection allows for a more comprehensive assessment of the models' ability to reason across different medical contexts and complexities. Additionally, for conditions such as Type II diabetes and Vitamin A deficiency, the answers may seem quite straightforward, making them useful for assessing whether the models can correctly identify and reason over well-established medical knowledge. Whereas, for complex conditions such as Lung Cancer and Alzheimer's Disease, we can evaluate the models ability to reason through more intricate, multi-factorial diseases.

Our qualitative analysis showed several key findings regarding the models' reasoning capabilities across different diseases. Interestingly, none of the models generated questions for Alzheimer's Disease, which was unexpected given its significant global impact and strong presence in the Knowledge Graph. In contrast, the models seemed to reason well over diabetes-related questions, although it would often provide multiple correct answers, even when prompted for a single response. This could suggest an alignment with well-established medical knowledge in this domain. For cancer-related questions, the models tended to select the most straightforward and common answers, though the Knowledge Graph contained a broader mix of more complex phenotypes. This seems to indicate a preference for simplicity in model-generated reasoning, potentially overlooking more nuanced aspects of the disease.

When comparing open-source models against proprietary models, the qualitiative analysis shows that proprietary models generally performed better than open-source models in terms of providing structured and direct responses. proprietary models demonstrated a better adherence to the prompt constraints, whilst the open source models seem to show more explanatory or multi-component answers. For example, the HuatuoGPT-70B open source model, consistently responded with "thinking" before elaborating on its reasoning instead of directly providing a single answer for both 1-hop and 2 hop prediction as prompted. This suggests that the model prioritises explaining its reasoning than strictly following the prompt's format. In constrast, however, proprietary models such as GPT-4 more reliably adhered to the prompt constraints,. When prompted to give a single answer for one hop and two hop questions, GPT-4 consistently did so, and this was present across the closed-source GPT family, suggesting that these proprietary models may be better optimised for tasks requring direct and efficient responses. Among the open source models tested, LLaMA Ultra Medical, a medical open source LLM, tended to provide multiple answers when prompted for one single answer, and of those multiple answers, apart from Type II diabetes, most answers were incorrect.

Taking an look into responses related to diabetes, responses were quite mixed. For instance, HuatuoGPT-01-8B performed outside the constraints of the task, correctly predicting the answer before proceeding to provide its reasoning. On the other hand, LLaMA 8B Instruct struggled with both Hop 2 and Hop 1 predictions, failing to generate the correct responses. Similarly, LLaMA Ultra Medi-

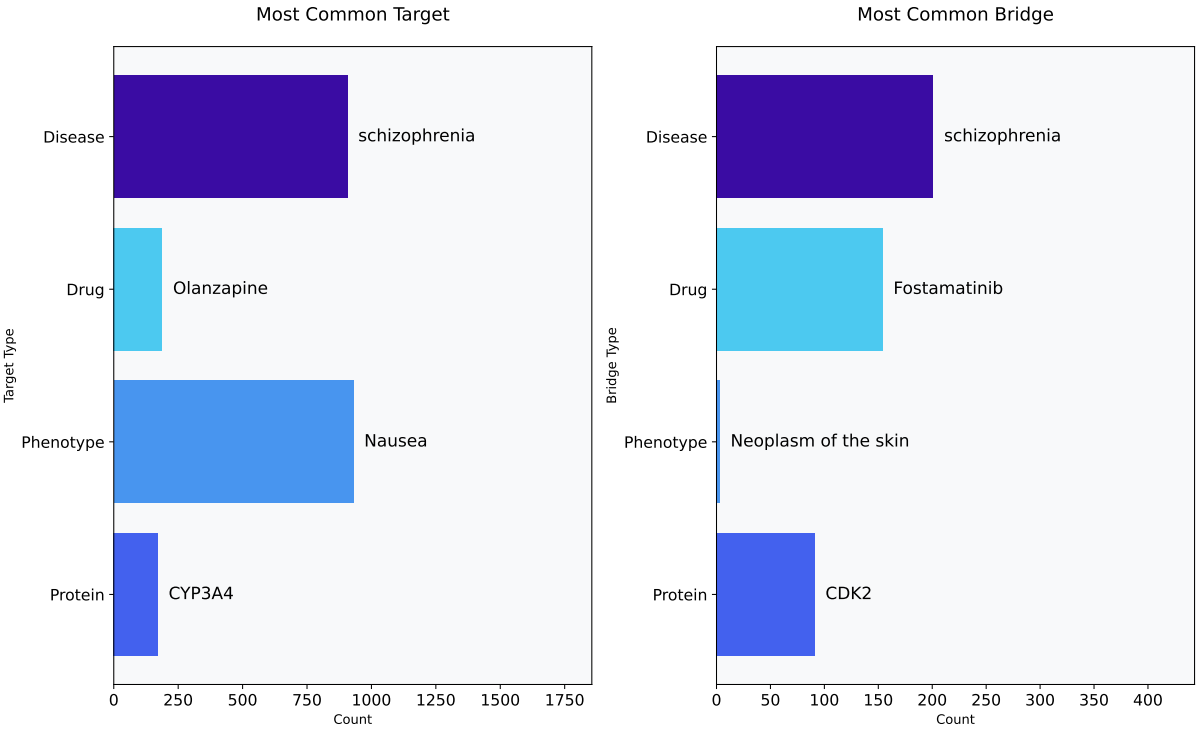

Figure 2: Common target and bridge entities for each node type in BioHopR.

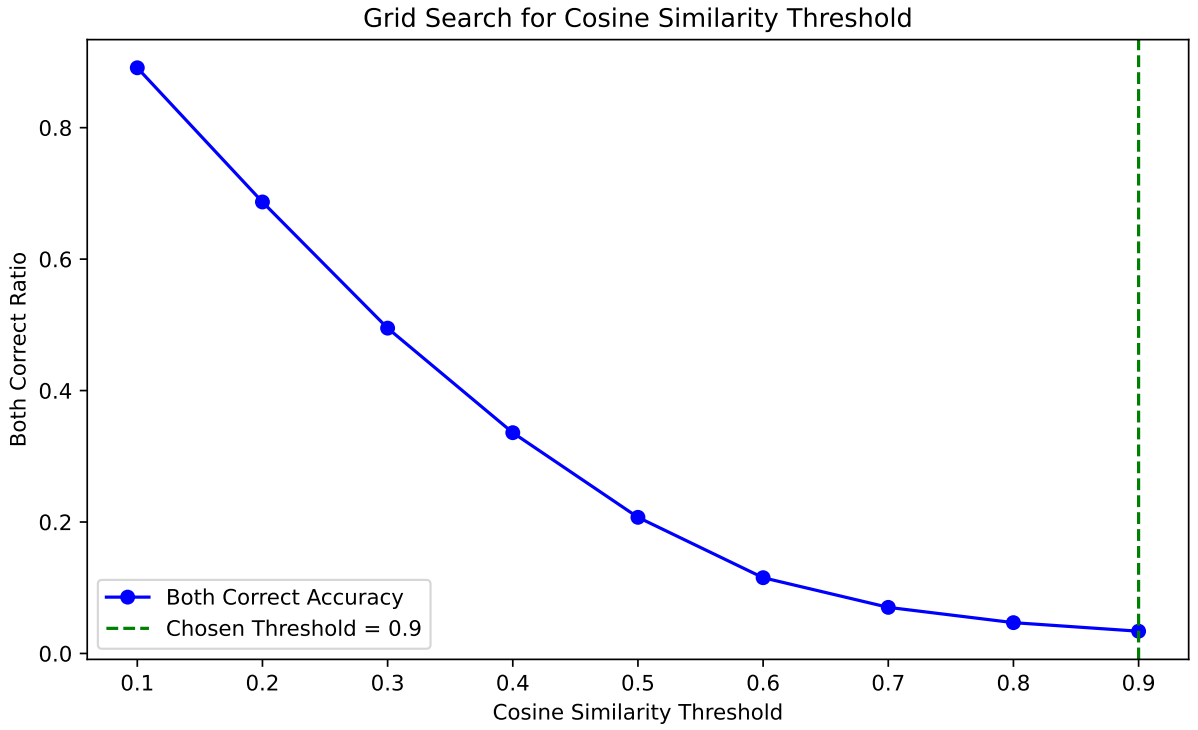

Figure 3: Grid search results showing the relationship between cosine similarity threshold and accuracy for "Both Correct" predictions. The chosen threshold of 0.9 is marked, reflecting the strict precision requirements in the biomedical domain.

cal did not fully adhere to the prompt's instructions — when asked to provide a single answer for Hop 2, it instead generated a list of multiple possible answers. While the listed responses were correct, this deviation indicates a challenge in following explicit task constraints. Moreover, for Hop 1, the model's response was incorrect, further highlighting inconsistencies in its performance. Interestingly, GPT-4 did not correctly predict the Hop 2 or Hop 1 answers in a strict sense. However, the model consistently included hepatotoxicity as a response—a condition that, while not explicitly listed among the correct answers, is still a relevant and justifiable finding. This pattern was observed across the GPT model family, suggesting that these models might apply broader medical reasoning even when their direct predictions do not align with predefined correct answers.

Schizophrenia, as seen in our figure, appeared frequently in the data. For GPT-4, in one-hop predictions, the model frequently guessed Clozapine as a treatment for schizophrenia. While this answer is medically correct, it was not explicitly part of the datasets predefined answer set. This suggests that the model is leveraging broader clinical knowledge rather than strictly adhering to the dataset's constraints. This trend was also consistent across the other GPT-family models, GPT-4o mini. However, for the o3 models, the one-hop predictions were correct and within our predefined list of answers.