BioMol-MQA: A Multi-Modal Question Answering Dataset For LLM Reasoning Over Bio-Molecular Interactions

Anonymous ACL submission

Abstract

Retrieval augmented generation (RAG) has shown great power in improving Large Language Models (LLMs). However, most existing RAG-based LLMs are dedicated to retrieving single modality information, mainly text; while for many real-world problems, such as healthcare, information relevant to queries can manifest in various modalities such as knowledge graph, text (clinical notes), and complex molecular structure. Thus, being able to retrieve relevant multi-modality domain-specific information, and reason and synthesize diverse knowledge to generate an accurate response is important. To address the gap, we present BioMol-MQA, a new question-answering (QA) dataset on *polypharmacy*, which is composed of two parts (i) a multimodal knowledge graph (KG) with text and molecular structure for information retrieval; and (ii) challenging questions that designed to test LLM capabilities in retrieving and reasoning over multimodal KG to answer questions. Our benchmarks indicate that existing LLMs struggle to answer these questions and do well only when given the necessary background data, signaling the necessity for strong RAG frameworks.

1 Introduction

LLMs (Zhao et al., 2023) have shown great performance in various tasks (Srivastava et al., 2023), such as text summarization (Zhang et al., 2024a) and question answering (QA) (Singhal et al., 2025). Their success has attracted people to adopt them for daily use, e.g., to address medical concerns (Mendel et al., 2025; Yun and Bickmore, 2025). Despite their strong abilities, they face issues such as hallucination (Huang et al., 2025), knowledge cutoff (Cheng et al., 2024) and lacking domain knowledge (Yang et al., 2023; Gu et al., 2025). This hinders their adoption in high-stakes scenarios as healthcare where incorrect responses can mislead users putting their lives at risk.

DRUG 1: Ketorolac
DRUG 1 BACKGROUND: ... is a potent nonsteroidal anti-inflammatory drug (NSAID) indicated for the management of moderate to severe nociceptive pain ... resulting in the attenuation of prostaglandin synthesis ...
DRUG 2: Oxaprozin
DRUG 2 BACKGROUND: ... is a nonsteroidal anti-inflammatory drug (NSAID), ... used to relieve joint pain associated with osteoarthritis and rheumatoid arthritis. Chemically, it is a propionic acid derivative...
DRUG-DRUG INTERACTION: Ketorolac-gastric inflammation-Oxaprozin

Q: Which medication, used for moderate to severe nociceptive pain and known for robust inhibition of prostaglandin synthesis, may contribute to gastric inflammation when used in combination with a propionic acid derivative NSAID indicated for arthritis?
A: Ketorolac

Figure 1: Example from BioMol-MQA. Colors highlight elements from each information source (Blue = Graph, pink|orange = Text) that a model must connect to answer questions. Here, an LLM needs to first identify the correct NSAID (Oxaprozin), determine its outgoing edges in the graph that have a gastric inflammation label, and figure out which drug on the other end is used to manage nociceptive pain. More examples in App. I.

A popular way to address these issues is Retrieval Augmented Generation (RAG) (Lewis et al., 2020) - retrieving query relevant context to ground an LLMs response on it for for factual accuracy. Various RAG methods (Fan et al., 2024; Gao et al., 2023) have emerged over time. However, they mainly focus on retrieving information from a single modality (Jin et al., 2024), e.g., text (Trivedi et al., 2023), knowledge graph (Edge et al., 2024), or image (Shalev-Arkushin et al., 2025) only. Although there have been some attempts for multi-modal RAG (Abootorabi et al., 2025), i.e., retrieval from more than one modality, they mainly focus on general knowledge domains (Jin et al., 2024; Abootorabi et al., 2025) that do not require expertise in any particular area. For many real-world problems, e.g., healthcare, information relevant to medical queries can manifest in various modalities such as KGs, text (clinical notes), and complex

molecular structure from crystallography. Thus, being able to retrieve the relevant domain information from multiple modalities, and reason over diverse knowledge to generate an accurate response is crucial for LLMs. However, the work on understanding and developing LLMs with such ability is limited (AlSaad et al., 2024). One impediment is the lack of datasets supporting this line of research.

To fill this gap, we develop a multi-modal retrieval and reasoning dataset named BioMol-MQA for QA on *polypharmacy*. Polypharmacy (Halli-Tierney et al., 2019; Masnoon et al., 2017), the concurrent use of multiple medications to address ailments, is a serious issue in healthcare where the goal is to combat multiple conditions with a combination of drugs. However, unknown drug-drug interactions may exacerbate rather than improve one's health. Given the seriousness of the issue and the extensive domain knowledge required for answering related questions, polypharmacy is a perfect testbed to gauge an LLM's multi-modal reasoning abilities. Here, an LLM must utilize background information on drugs, their interaction partners, and molecular-level details to determine the severity of a drug combination. Additionally, it is known that most pharmaceutical drugs target proteins (Vicidomini and Roviello, 2023; Atlas, 2025). Having access to protein-level data will help in better resolving polypharmacy queries.

BioMol-MQA is composed of two parts: (i) a multi-modal (KG) for information retrieval; and (ii) complex medical questions that require multi-modal RAG and LLM reasoning for QA. The multi-modal KG provides complex and comprehensive domain knowledge in three modalities: a KG of drug-drug and drug-protein interactions, free-text on each entity providing background knowledge, and molecular structure data on drugs, encoded as SMILES strings (§3.1.3). Our questions are constructed by integrating information from the above three diverse sources. To answer, an LLM must retrieve information from the three modalities to provide an accurate response. To facilitate evaluation and model fine-tuning, each question has a ground truth answer and resources from which the problem was constructed (§3.4). An example is shown in Figure 1. We formally define the task as,

Task Definition *Given a multi-modal data structure $\mathcal{D}(\mathcal{G}, \mathcal{T}, \mathcal{S})$ consisting of a knowledge graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ ($\mathcal{N}$ = Nodes (drugs and proteins) and $\mathcal{E}$ = Edges connecting the nodes), a text corpus on the knowledge graph entities $\mathcal{T}$ and molecular structures of drugs as SMILES strings, $\mathcal{S}$, an LLM must utilize at least two modality combinations ($\mathcal{G} + \mathcal{T}$ or $\mathcal{G} + \mathcal{S}$) to solve a query $\mathcal{Q}$, whose answer is a node in the graph. Mathematically, the task can be described as a function $\mathcal{F} : (\mathcal{D}, \mathcal{Q}) \rightarrow \mathcal{N}$.*

Experimental results on BioMol-MQA (§4) show that current LLMs, on their own, are inept at solving these questions. However, when provided with the relevant context, their performance spikes, indicating the necessity to ground reasoning on information via RAG. BioMol-MQA also provides data for benchmarking multi-modal retrievers. Additionally, it has a dedicated training, validation and test split (§3.4) that can be utilized for fine-tuning models to jointly learn question and modality (graph/text/SMILES) representations useful for downstream tasks as graph link (determining existence of edges between nodes) (Li et al., 2023) or molecular property prediction (Guo et al., 2023).

Our **main contributions** are: (i) We propose BioMol-MQA, a new question answering dataset for retrieval and reasoning over multi-modal contexts; (ii) We propose a synthetic data generation pipeline for augmenting knowledge graphs and creating questions that integrate diverse modalities; (iii) We show the limitations of current retriever models when querying our knowledge source and frontier LLMs in reasoning in complex scenarios as ours.

2 Related Work

RAG In Medicine Although most RAG pipelines are tested on general domain data (Jin et al., 2024; Abootorabi et al., 2025), there are studies exploring RAG in healthcare. Lozano et al. (2023) builds an application for QA over PubMed (Wikipedia contributors, 2025) articles. Wang et al. (2023) perform molecule synthesis by retrieving similar samples from a labeled database. However, they follow a *unimodal* setup, i.e., retrieve information from a single-source. Although Wang et al. (2024c) perform QA over a multi-modal biomedical KG, the questions themselves do not integrate each source, essentially collapsing into unimodal retrieval.

RAG Datasets Current benchmarks for RAG, such as RAGBench (Friel et al., 2024) and ChatRAG Bench (Liu et al., 2024c), are a combination of existing QA datasets. The original QA datasets are augmented with corpora for RAG testing. That said, we do find two new datasets that directly support RAG, viz., CRAG (Yang et al., 2024) and STaRK (Wu et al., 2024b). CRAG only discusses

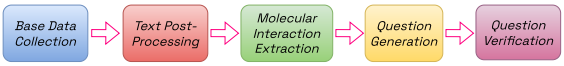

Figure 2: Our dataset development pipeline.

unimodal questions related to the general domain. STaRK combines questions from graph and text modalities to support multi-modal graph retrieval. However, for each KG triple (entity-relation-entity), they only consider the background-text on one entity and do not *strongly* integrate relationships (App. E). This limits question complexity and coverage.

Molecular QA Recently, QA on molecules has seen growth- `MoleculeQA` (Lu et al., 2024) and `PubChemQA` (Luo et al., 2024). However, questions in both are *template-based*. `PubChemQA` asks *describe this molecule* while `MoleculeQA` has types such as *Which kind of compound does this molecule belong to?* This heavily limits sample diversity. Further, neither considers *relationships* between molecules and, are restricted to one modality by ignoring background knowledge of the molecules.

`BioMol-MQA` addresses these limitations by effectively combining multiple modalities, utilizing entity relationships and creating questions that capture all of this. App. E has a detailed related work.

3 Dataset Construction

The overall pipeline of developing our dataset is summarized in Figure 2. First, we *assemble our base data* (§3.1) where we label our KG and augment its nodes with background text and molecular structures. This transforms the base KG into a multi-modal data structure. Next, we post-process (§3.2) our corpus to enhance question complexity. To address the absence of molecular interactions in the KG, we use an LLM to augment the graph (§3.3), adding another source for question generation. Next, we create our dataset of LLM-derived questions incorporating each modality (§3.4). Finally, we assess the quality of our dataset (§3.5) with the help of a human and automatic evaluator.

3.1 Base Data Acquisition (Stage I)

To fully test the ability of existing systems for retrieving and reasoning multi-modal information to answer complex medical queries, our dataset offers three modalities that they must utilize for answering: (i) a **knowledge-graph** of drugs and proteins, where each node is a drug or protein and each edge is an interaction between two entities; (ii) a **free-text** associated with each node in the graph, which provides background information such as uses, behavior, etc.; and (iii) the **molecular structure** of drugs represented by a non-natural language string called SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988; Zhang et al., 2025), which gives insight into the components potentially responsible for participating in various bio-molecular interactions.

3.1.1 Graph Modality

Our base knowledge graph is from Zitnik et al. (2018). The graph consists of two types of nodes, drugs and proteins, and three types of edges, where each edge is an interaction between the corresponding entities. In total, there are 645 unique drugs, 19K proteins, ~4M drug-drug interactions (DDI), ~130K drug-protein interactions (DPI), and ~715K protein-protein interactions (PPI). Only D-D edges are labeled (~1.3K unique) by their polypharmacy side-effects mined from various medical databases.

Component	Count
Drugs	494
Proteins	198
Drug-Drug Edges	18585
Drug-Protein Edges	314
Drugs with Text	179
Proteins with Text	51
Drugs with SMILES	494

Table 1: Graph Metrics

We do not create questions involving each edge (interaction). This is because, not all nodes (drugs/proteins) have mineable text-data (§3.1.2) which we require for question generation (§3.4). Thus, to maintain decent coverage while ignoring unusable nodes, we sample (App. A) a subgraph with statistics as shown in Table 1.

Resolving Entity Names. The base knowledge graph provides entity names in a coded format. Drugs are represented by their `CID` (Compound Identifier) such as `CID000002088` (*alendronic acid*), while proteins are described by their NCBI (National Center for Biotechnology Information) Entrez database entry (of Medicine, 2025) such as `3351` (`HTR1B`, gene name corresponding to the ID, but mappable to its protein). These are identifiers used in biomedical research. However, as we are building a natural language QA dataset, we cannot refer to entities by their IDs as they are only meaningful in the context of their respective databases and not scientific discourse. For example, `3386` (*Fluoxetine*) only makes sense when connected to the PubChem database (Kim et al., 2025) and typically not found in research studies. Moreover, we need to identify the names of those drugs/proteins for searching literature (§3.1.2) as using IDs yields no documents. Specifically, we obtain entity names by querying PubChem/STRING (Szklarczyk et al., 2023) database, depending on

which one can provide a generic name. App. D provides extensive details on name resolution.

Labeling Drug-Protein Interactions (DPI). The base knowledge graph does not contain edge labels for DPIs. Without knowing what relationship exists between a drug and a protein, we cannot create questions about them. To address this, we use `STITCH` (Search Tool For Interactions Of Chemicals) (Kuhn et al., 2007), a database integrating information on small molecules (drugs) and their associations with proteins. For each D-P pair (c.f. Table 1), we query STITCH to retrieve their interaction, which is provided in two fields; *mode* (nature of the interaction; *binding*, etc.) and *action* (effect of the interaction; *activation* or *inhibition*). We combine both fields to label a DPI, which results in two edge types, i.e., *binding and activation* and *binding and inhibition*. Further details in App. C.

3.1.2 Text Modality

The second modality is text. To associate each node with a document,we use Wikipedia summaries. We consider sourcing abstracts from PubMed (Wikipedia contributors, 2025). However, they are usually clinical studies about a *specific* aspect of a drug/protein as opposed to broader knowledge. Additionally, they are short and combining many leads to an incoherent document for processing. Thus, we use Wikipedia which provides sufficient and succinct background for an entity.

We query Wikipedia for each entity in our graph. As explained in App. D.1, while many drugs are registered in medical databases, there is no guarantee that we can obtain documented information about them, apart from basic properties, due to most drugs failing clinical trials (Sun, 2022). As such, we focus on those drugs that either have a generic or commercial name identified in §3.1.1 to make it easier to obtain background data. This excludes drugs such as `(+)-Vinblastine` (stereochemistry information) giving a total of 198 drugs and 51 proteins with Wikipedia entries. These entities yield 500 DDIs and 84 DPIs (Table 3).

3.1.3 SMILES

Our final modality is the molecular structures of drugs. Drugs are ultimately compounds with distinct behaviors and structures. As such, there is valuable information to be gleaned by studying their molecular composition, such as reaction sites and functional groups, which motivates us to include drug structure as our third modality. SMILES (project, 2020) is a popular notation used to represent chemical compounds in electronic databases. It provides a set of rules to define molecular structure in a *non-natural* string. For example, `C=C` indicates `Ethene` (CH_2CH_2) and = refers to a double bond between the carbon atoms. Using these rules, drugs and compounds are standardized. Proteins do not have a SMILES representation as they are not considered "small molecules" like drugs (Lin et al., 2019). Thus, we can only use the molecular structure of drugs in our dataset. We obtain SMILES for each drug in the KG by querying PubChem (Kim et al., 2025), a database of chemical compounds. Further SMILES details are provided in App. B.

3.2 Text Post-Processing (Stage II)

	Before PP	After PP
Tokens	300	233
Readability	15	28

Table 2: Post-processing (PP) outcomes. Numbers represents the average over the documents.

Most Wikipedia summaries are condensed and have the relevant biomedical data we need. However, a summary may also include irrelevant information such as market statistics and historical data. Additionally, many summaries are verbose and can inundate an LLM's context window (Liu et al., 2024b). To address these issues, we utilize GPT's paraphrasing abilities (Hassanipour et al., 2024). First, we identify documents having more than 200 tokens (empirically). This excludes all protein texts as they have an average of 114 tokens, while for drugs it is 212. Next, we prompt (Fig. A2) GPT-4o (Hurst et al., 2024) to rewrite the source text by ignoring historical data and using domain-specific jargon suitable for an expert audience. There are two reasons for doing this: i) to enhance question complexity based on the rephrased text; and ii) to reduce their chance of *cheating* by using direct references from Wikipedia, as most LLMs have extensive training on it (Johnson et al., 2024).

We assess the result of this process in two ways, i.e., *average tokens* for measuring conciseness and *readability* scores for measuring text complexity. For readability, we use `Gunning-Fog` Index (Gunning, 1969), which indicates the level of expertise required to comprehend a given text. Higher values imply increased complexity suitable for audiences with formal training in a discipline. Results from the process are in Table 2. As we see, the enhanced texts are more condensed with readability almost doubled. This achieves our goal of having a semantically dense corpus addressing the above issues.

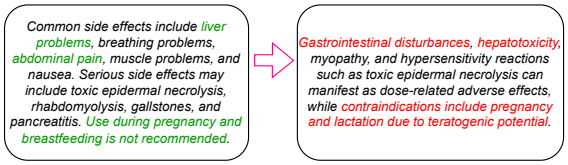

Figure 3: Simple phrases as *liver problems* are replaced by technical jargon as *hepatotoxicity* by post-processing.

Fig. 3 shows the effect on *Fenofibrate*'s (drug) text.

3.3 Molecular Interactions (Stage III)

The relationships described in the initial knowledge graph are at the *physio/biological level*, i.e., physical manifestations of taking two drugs (e.g., *nausea*, *headaches* etc.) or influence of a drug on protein activity (*binding*, *activation*, etc.). However, there are interactions that exist at the *molecular* level which describe potential chemical associations between the atoms of molecules. Drugs are small molecules that also have such interactions between them. Unfortunately, molecular relationships are not provided by the base graph and ignoring them would lead to loosing insights obtained from a crucial aspect of drug interaction.

To remedy the absence of molecular interactions, we leverage GPT-4o's knowledge of chemistry. As shown by prior studies (Guo et al., 2023; Hatakeyama-Sato et al., 2023), LLMs have strong capabilities in reasoning through chemistry tasks, including molecular property prediction from SMILES analysis (Guo et al., 2023). This provides credence in its ability for molecular interaction extraction. Given the molecular structure (SMILES), GPT-4o is asked to describe potential associations between the atoms two drugs. The interpretable nature of LLMs provides better insight into their reasoning process for such a complex task.

We have GPT output (Fig. A3) four fields, i) *interaction name* - a label for the potential interaction ii) *mechanism* - why it occurs, iii) *evidence* - cues from the SMILES to lend credence to its reasoning and iv) *severity* - how strong the interaction is (low/medium/high), if it can be identified. Of the 500 drug-drug pairs that we use for question-generation (c.f. 3.1.2), GPT is able to extract molecular relationships for 499 pairs. Investigating the generations reveal five types of relationships: `Hydrogen Bonding`, π-π `Stacking`, `Steric Clashes`, `Ionic` and `Electrostatic Interaction`. The introduction of these interactions transforms the base graph into a *multi-graph*, i.e., a graph with multiple edges between two nodes, adding a further layer of nuance. Fig. 4 depicts a portion of the final graph.

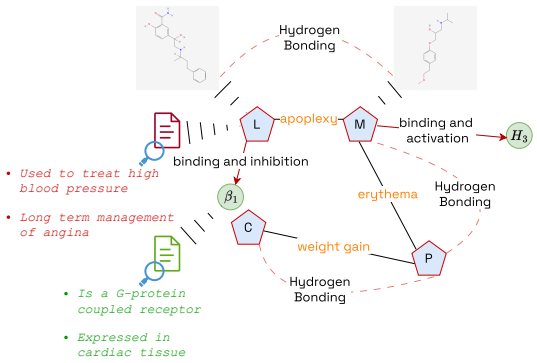

Figure 4: The final knowledge-graph depicting each modality, edge direction and augmented edges (molecular interactions). Entities in blue are drugs, and those in green are proteins. Dashed edges indicate molecular interactions, and solid edges are DDI/DPI, depending on the type of entities. L = *Labetalol*, M = *Metoprolol*, β_1 = *Beta-1 adrenergic receptor*, H_3 = *Histamine H3 receptor*, P = *Pirbuterol*, C = *Carvedilol*

3.4 Question Generation (Stage IV)

LLMs have demonstrated great performance in rewriting or generating desired questions that follow a given prompt and context (Abdullin et al., 2023; Wu et al., 2024a; Hematian Hemati and Beigy, 2024; Liu et al., 2024a). Leveraging these capabilities, we *prompt* GPT-4.1 (OpenAI, 2025a) for our question generation (Fig. A5). For a knowledge graph triple (entity-relation-entity), GPT is either given background text on each entity or their SMILES (for drugs). It is then asked to create a question integrating the relational information with the background data such that the answer is one of the respective KG-nodes (Balepur et al., 2025). The idea is that, **to answer the questions, a model must utilize multiple information sources (graph/text/molecular structure) to provide an accurate response**. Each question has a single gold-truth answer (one of the nodes associated with the respective KG triple) and associated background data needed to answer it.

Our questions are not multiple-choice, i.e., we do not provide models with options to choose from. This is a design choice made to reflect real-world information-retrieval (IR) settings (Abbasiantaeb and Momtazi, 2021; Biancofiore et al., 2024) wherein users query databases without associated choices. Questions based on DDIs are of two types according to the relationship (physio/biological and molecular), while there is only one type of question

Entity Pair	Interaction Type	1-hop	2-hop	3-hop	Total
Drug - Protein	DPI	84	100	100	284
Drug - Drug	Physiological (Bio)	500	100	100	700
	Molecular	499	100	100	699

Table 3: Question distribution in the dataset.

for a DPI. To craft questions, GPT is always provided with the knowledge-graph interaction triples between the respective entities. Additionally, it receives background text on each entity (for Bio-based DDIs and DPIs) or SMILES strings (for molecular DDIs). The answer for a DDI is always a drug, while for a DPI is the protein, as we wanted to capture the directional nature of the interaction.

The above questions utilize properties of two connected entities, which are *single-hop* questions. To add a further layer of complexity, we also create *multi-hop* reasoning questions, i.e., those involving two entities connected by a path longer than 1. For this, we randomly sample 100 two- and three-hop DDI and DPI edges. The overall distribution of questions is in Table 3. We divide the questions in a stratified 80-10-10 ratio into training/validation/test, resulting in 1346, 169, and 168 samples in the respective splits. For each question, we have the ground truth answer and the ground truth data (KG triples/entity background text/SMILES) used to construct the question.

We formulate two key requirements for crafting questions: (i) **Entity names must not appear verbatim in the question** - To push a model to understand the context, connect it to the knowledge base and then infer the entity name as opposed to directly learning the entities involved. This aligns with our goal of probing the multi-modal information processing abilities of LLMs; and (ii) **Questions should always test for the relationship between entities rather than isolated facts on each** - As our objective is to incorporate relational (graph) data and frame questions around it, instead of developing straightforward factoid questions.

For criteria one, we check for the presence of the entities used to create the respective question. Overall, we do not find any major leakage, thus affirming our first criteria. The only exception is that 15% of DPI questions mention the protein name. We excuse the LLM for doing this as proteins, overall, do not have much background data for it to utilize. Except that, none of the questions mention drug names or leak their SMILES strings, showing that the model followed our instructions well.

For criteria two, we notice that most questions have identifiable phrases to indicate a relationship between the entities. For example *combined with*, *associated with*, etc. We use the proportion of questions having these phrases as a rough proxy for our relationship criteria. This yields a ~73% hit rate, indicating that the majority of questions follow our instructions. As it is difficult to automatically evaluate the other questions due to variations in phrasing, holistic examination indicate that all questions have a relational component from the KG.

Metric	Value
Question Length (Tokens)	66.6
Type-to-Token ratio (TTR)	0.84
Shannon Entropy	5.67
Dependency Parse Tree Depth	10.86

Table 4: Question Statistics.

Quantitative metrics for analysis are: Question Length, Type-to-Token Ratio (number of unique words/total words), Entropy (how much information is conveyed by the text) and dependency parse tree depth (a structure to describe grammatical roles and relationships between words). App. M has details on each metric. Scores in Table 4 reflect the linguistic and semantic complexities of our dataset.

3.5 Question Verification (Stage V)

We assess question quality by (i) **Automatic Verification**. Using LLMs to assess data quality (`LLM-as-a-judge`) (Gu et al., 2024; Li et al., 2024a,b) has emerged as a promising proxy for human evaluation. By describing a set of grading criteria (rubric), frontier LLMs (Achiam et al., 2023; Hurst et al., 2024) can provide assessments aligning with real annotators (Gu et al., 2024; Li et al., 2024a,b). However, using the same LLM for generation and evaluation can introduce biases (Panickssery et al., 2024). Thus, we use `Claude-3.7 Sonnet` (Anthropic, 2025) for evaluation; (ii) **Human Evaluation**. Despite the proficiency of `LLM-as-a-judge`, they are prone to issues such as preferring verbose outputs and prompt sensitivity (Koo et al., 2024). Thus, having a human-in-the-loop + judge-LLM is effective for gauging data quality. One of our co-authors, a PhD student in Bioinformatics provides the domain-knowledge required to evaluate the questions.

Metric	Human	Automatic (LLM)
Clarity	1.22	1.75
Coverage	1.79	1.97
Assumptions	1.79	1.98
Inference	1.79	1.91

Table 5: Human vs. Automatic Question Eval.

We randomly sample 100 QA pairs for the LLM and human annotator to provide feedback. Both are given a rubric (Fig. A4) of 4 custom metrics: *Clarity* (technical

prowess needed to understand the question); *Coverage* (proportion of modalities utilized); *Assumptions* (including information beyond what was supplied (hallucination)); *Inference* (can the answer be derived by studying the provided data (Jullien et al., 2024)). Each metric is scored as 0, 1 or 2, depending on the guidelines (Fig. A4). Averaged scores for each metric for human and LLM evaluation is in Table 5. As we see, there exists an overall agreement between our human evaluator and the LLM. The only measure where the two differ is in *Clarity*, with the LLM recognizing that the questions on average are harder (higher scores = greater complexity). This makes sense as `Claude` is a generalist model (Guan et al., 2024) while our human annotator is a well-read medical student, conversant with the language/content of the questions.

4 Experiments

We perform two categories of experiments; *LLM-reasoning* and *retrieval*-based generation.

4.1 LLM Reasoning Capability

Here, we investigate LLMs' reasoning capabilities (i) zero-shot, i.e., without `RAG` - lower bound and (ii) when they are given the gold data (KG-triples, background text/SMILES) - upper bound.

LLMs Used We benchmark seven models on the test split of our dataset, which include: (i) Two closed-source LLMs, (Anthropic, 2025) and `o4-mini` (OpenAI, 2025b); and (ii) five open-source LLMs, `TxGemma` (Wang et al., 2025a) (trained to understand information on various areas of drug development), `DeepSeek-R1` (Guo et al., 2025), `Llama 3.3` (Meta, 2024), `Qwen3` (Team, 2025), and `Mistral` (Jiang et al., 2023a). Each model is first asked to provide their thought-process to arrive at the answer and then the answer itself. This strategy of prompting has been found to improve LLM performance according to recent studies (Wang et al., 2024a; Yugeswardeenoo et al., 2024). We use *Together AI* (AI, 2025) for evaluating the best variants of open-source LLMs.

Metrics LLM answers are evaluated using (i) Lexical EM (exact match): 0 (miss)/1 (hit) measure of string equality, (ii) Lexical F1: Balanced proportion of token overlap between the predicted and reference strings and (iii) `BERTScore` F1 (Zhang* et al., 2020): An embedding-based similarity measure between two strings. By using these three metrics covering different aspectswe can have a better understanding of LLM reasoning capabilities and our dataset quality.

Analysis Table 6 shows the results. We observe: **(i)** The average zero-shot EM/F1/BERTScore (across all models) is 0.22, 0.28, and 0.77, respectively. This clearly shows that existing LLMs are not yet equipped to handle these types of questions by themselves as they lack the necessary domain knowledge; **(ii)** The average EM/F1/BERTScore (across all models) for the upper-bound test is 0.62, 0.67, and 0.83, respectively. This indicates two things, **(1)** *the questions in our dataset, while challenging, are answerable and fair as supported by these scores*, which demonstrate the quality of our data; **(2)** the importance of grounding answers through RAG, i.e., *LLMs are capable of handling these questions, provided they are given the necessary background information*, which further show the necessity of our multimodal KG. For example, models such as Claude and DeepSeek almost triple their EM while Mistral's F1 improves five times.

4.2 Retriever Performance / Multimodal RAG

Here, we examine how well current retrievers can query our knowledge sources. We test three types of retrievers, text-only, graph-only and a combination of both as our multi-modal baseline which simply combines the best text/graph retriever.

Retrievers Used For text-only retrieval, we test *sparse*: BM25 (Schütze et al., 2008) and *dense*: [MedCPT (Jin et al., 2023), DPR (Karpukhin et al., 2020), MolLM (Tang et al., 2024), `text-embedding-3-large` (TE3L) (OpenAI, 2024)] models. BM25 is a competitive baseline (Luo et al., 2023), occasionally surpassing dense models (Luo et al., 2023). MedCPT, DPR and MolLM are BERT-style (Devlin et al., 2019) encoders, trained for different purposes such as molecular data understanding (Tang et al., 2024) and QA (Karpukhin et al., 2020). TE3L is the only proprietary model. Details on each is in App. K.

For our graph-only retriever, we implement a simple baseline using the `Neo4J` database (Neo4j, 2025). It works by retrieving KG triples semantically similar to the query. The query and triples are encoded using `all-MiniLM-L6-v2` embeddings (sentence transformers, 2021) (other embedding models performed worse). Although a simple baseline, we find it to be reasonably effective. That said, given the complexity of the graph, we only use the set of triples for question creation (§3.4) for

Model	Closed Source LLMs		Open Source LLMs				
Approach	o4-mini	Claude 3.7 Sonnet	DeepSeek R1	LLama 3.3	Qwen 3	TxGemma	Mistral
Zero-Shot	**(0.32, 0.42, 0.79)**	(0.30, 0.36, 0.78)	**(0.33, 0.40, 0.87)**	(0.27, 0.34, 0.85)	(0.22, 0.25, 0.51)	(0.07, 0.16, 0.82)	(0.02, 0.14, 0.81)
Upper Bound	(0.88, 0.88, 0.94)	**(0.89, 0.89, 0.94)**	**(0.90, 0.90, 0.98)**	(0.71, 0.71, 0.93)	(0.80, 0.80, 0.88)	(0.76, 0.76, 0.94)	(0.74, 0.76, 0.95)

Table 6: Reasoning benchmark scores (without RAG). **Bold** represents the best performing model in each category for the corresponding test. Each tuple is (lexical EM, lexical F1, BERTScore F1)

Retriever	Hit@5	Hit@10	Hit@15	Recall@5	MRR
BM25	**0.2/0.58**	**0.28/0.58**	**0.3/0.58**	**0.4**	**0.52**
DPR	0/0.14	0.01/0.23	0.06/0.27	0.06	0.07
MedCPT	0.10/0.49	0.15/0.56	0.21/0.57	0.27	0.45
MolLM	0.01/0.38	0.03/0.47	0.04/0.51	0.17	0.24
TE3L	0.15/0.57	0.24/0.58	0.28/0.58	0.35	0.49
Neo4j*	0.15/0.19	0.19/0.26	0.2/0.28	0.05	0.13
Hybrid	0.06/**0.60**	0.13/**0.60**	0.14/**0.61**	0.32	**0.52**

Table 7: Retriever performances. The first 5 rows are text-only retrievers. SMILES information is encoded as text and falls under text-only retrieval. *Neo4j was run on the subset of triples used for question generation due to poor performance on the entire graph. Hits are provided as hard/soft. Hybrid = [BM25 + Neo4j]

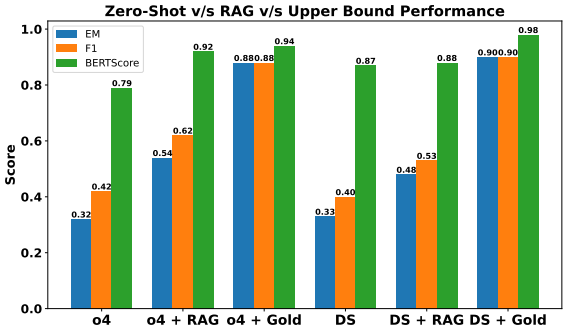

Figure 5: Impact of RAG on o4-mini and DeepSeek R1 (DS). For comparison, their zero-shot and upper bound (with the gold data) performances are provided.

retrieval. We provide additional results in App. L.

Metrics Retrieval accuracy is measured using three metrics, i) Hits@k - (0/1) measure to see if *any* (soft hits) or *all* (hard hits) relevant items are among the top-k retrieved results; ii) Recall@5 - The proportion of correct items among the top-5 retrieved results iii) MRR (Mean Reciprocal Rank) - Inverse index of the first relevant result.

Analysis Results from benchmarking retrievers are in Table 7. We observe: **(i)** For text-only retrievers, a simple BM25 baseline outperforms *each* dense embedding model. This is supported by studies (Chen et al., 2022b; Sciavolino et al., 2021) which show how lexical models such as BM25 can generalize to different domains, occasionally outperforming dense embeddings. **(ii)** For our graph-only retriever, we notice a similar trend, where a basic database retriever (Neo4j) outperforms trained dense models (§App. L). As we explain in §App. L, training GNNs on our graph *directly* is difficult due to its small size (KG's typically have millions of nodes and edges (Ji et al., 2021)) and complexity. Thus, using simpler baselines (Neo4j) is preferable in settings as ours; and **(iii)** Our hybrid retriever (BM25+Neo4j) shows a balanced performance between the text-only and graph-only modes. Although the overall hard hit goes down, soft hit coverage is quite strong, outperforming each modality individually. This indicates the benefit of having even a simple multi-modal retrieval baseline.

Multimodal RAG Performance To test how well our models perform using RAG, we use our hybrid retriever (BM25/Text + Neo4j/Graph) with the two best zero-shot models, i.e., `o4-mini` and `DeepSeek-R1` (additional experiments in App. J). This experiment not only establishes the importance of grounding LLM responses on appropriate background knowledge but also, the limitations of existing retrievers in accessing complex data sources such as ours. Results from this test are provided in Fig. 5. The scores highlight three things, i) the importance of using even a basic retriever to provide domain-knowledge to LLMs, ii) poor performance of current retrievers in processing and providing multi-modal data to LLMs and, iii) the gap remaining between RAG-based reasoning and an LLMs true potential.

5 Conclusion and Future Work

We present `BioMol-MQA`, a multi-modal reasoning and retrieval benchmark for polypharmacy queries. Through our experiments, we show the limitations of existing retrievers and models in dealing with multi-modal contexts as ours. With our dataset, the goal is to provide a testbed for evaluating models for handling complex, real-world medical queries. We will continue to grow the dataset, such as including *unanswerable* questions, exploring *other modalities* as protein structures (long-chain amino acid sequence (Lopez and Mohiuddin, 2024)) for testing the length limits of LLMs (Wang et al., 2024b), scaling samples and using multi-agent workflows (Mitra et al., 2024) to enhance question diversity.

Limitations

We identify three limitations in our dataset; (i) Although multi-modal, only drugs in our dataset benefit from two modalities, i.e., SMILES and text. Proteins only have background text. This limits the type of questions we can craft for it. As we have mentioned in the conclusion, our next version will incorporate additional protein data; (ii) There are more drugs with SMILES than text in the knowledge graph, which creates an imbalance between the SMILES set and the text corpus in terms of size. We explain this in App. D.1 and §3.1.2 due to the nature of drug clinical trials; (iii) As our questions are very specific to the healthcare domain, they can only be qualitatively evaluated by domain practitioners. This potentially limits large-scale evaluation due to associated costs. Unfortunately, this is a reality in cross-domain studies involving biomedical topics where collaboration with experts is prohibitively restrictive.

References

Zahra Abbasiantaeb and Saeedeh Momtazi. 2021. Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6):e1412.

Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using LLM agents. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore. Association for Computational Linguistics.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Together AI. 2025. Together AI – The AI Acceleration Cloud - Fast Inference, Fine-Tuning & Training — together.ai. https://www.together.ai/. [Accessed 13-05-2025].

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

Anthropic. 2025. Claude 3.7 system card - Anthropic — docs.anthropic.com. [Accessed 29-04-2025].

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

The Human Protein Atlas. 2025. The human proteome in druggable - The Human Protein Atlas — proteinatlas.org. https://www.proteinatlas.org/humanproteome/tissue/druggable. [Accessed 11-05-2025].

Nishant Balepur, Feng Gu, Abhilasha Ravichander, Shi Feng, Jordan Lee Boyd-Graber, and Rachel Rudinger. 2025. Reverse question answering: Can an LLM write a question so hard (or bad) that it can't answer? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 44–64, Albuquerque, New Mexico. Association for Computational Linguistics.

Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. 2024. Interactive question answering systems: Literature review. *ACM Computing Surveys*, 56(9):1–38.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022a. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022b. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme.

2024. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era — blog.google. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. [Accessed 14-05-2025].

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Xiaodong Gu, Meng Chen, Yalan Lin, Yuhan Hu, Hongyu Zhang, Chengcheng Wan, Zhao Wei, Yong Xu, and Juhong Wang. 2025. On the effectiveness of large language models in domain-specific code generation. *ACM Transactions on Software Engineering and Methodology*, 34(3):1–22.

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 Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Haoxiang Guan, Jiyan He, Shuxin Zheng, En-Hong Chen, Weiming Zhang, and Nenghai Yu. 2024. Towards generalist prompting for large language models by mental models. *arXiv preprint arXiv:2402.18252*.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Anne D Halli-Tierney, Catherine Scarbrough, and Dana Carroll. 2019. Polypharmacy: evaluating risks and deprescribing. *American family physician*, 100(1):32–38.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Kimihiro Hasegawa, Wiradee Imrattanatrai, Zhi-Qi Cheng, Masaki Asada, Susan Holm, Yuran Wang, Ken Fukuda, and Teruko Mitamura. 2024. Promqa: Question answering dataset for multimodal procedural activity understanding. *arXiv preprint arXiv:2410.22211*.

Soheil Hassanipour, Sandeep Nayak, Ali Bozorgi, Mohammad-Hossein Keivanlou, Tirth Dave, Abdulhadi Alotaibi, Farahnaz Joukar, Parinaz Mellatdoust, Arash Bakhshi, Dona Kuriyakose, and 1 others. 2024. The ability of chatgpt in paraphrasing texts and reducing plagiarism: a descriptive analysis. *JMIR Medical Education*, 10(1):e53308.

Kan Hatakeyama-Sato, Naoki Yamane, Yasuhiko Igarashi, Yuta Nabae, and Teruaki Hayakawa. 2023. Prompt engineering of gpt-4 for chemical research: what can/cannot be done? *Science and Technology of Advanced Materials: Methods*, 3(1):2260300.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

Hamed Hematian Hemati and Hamid Beigy. 2024. Consistency training by synthetic question generation for conversational question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 630–639, Bangkok, Thailand. Association for Computational Linguistics.

Zhenyu Hou, Haozhan Li, Yukuo Cen, Jie Tang, and Yuxiao Dong. 2024. Graphalign: Pretraining one graph neural network on multiple graphs via feature alignment. *arXiv preprint arXiv:2406.02953*.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024a. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024b. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2).

Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. 2024. Wikimedia data for AI: a review of wikimedia datasets for NLP tasks and AI-assisted editing. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 91–101, Miami, Florida, USA. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Prentice Hall PTR. Online manuscript released January 12, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2025. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Stefanie Krause and Frieder Stolzenburg. 2024. From data to commonsense reasoning: the use of large language models for explainable ai. *arXiv preprint arXiv:2407.03778*.

Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. 2007. Stitch: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(suppl_1):D684–D688.

Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Juanhui Li, Harry Shomer, Haitao Mao, Shenglai Zeng, Yao Ma, Neil Shah, Jiliang Tang, and Dawei Yin. 2023. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. *Advances in Neural Information Processing Systems*, 36:3853–3866.

Tzyy-Shyang Lin, Connor W Coley, Hidenobu Mochigase, Haley K Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L Craig, Jeremiah A Johnson, Julia A Kalow, and 1 others. 2019. Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9):1523–1531.

Naiming Liu, Zichao Wang, and Richard Baraniuk. 2024a. Synthetic context generation for question generation. *arXiv preprint arXiv:2406.13188*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024c. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Advances in Neural Information Processing Systems*, 37:15416–15459.

Michael J Lopez and Shamim S Mohiuddin. 2024. Biochemistry, essential amino acids. In *StatPearls [Internet]*. StatPearls Publishing.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 8–23. World Scientific.

Shuai Lu, Nan Duan, Hojae Han, Daya Guo, Seung-won Hwang, and Alexey Svyatkovskiy. 2022. ReACC: A retrieval-augmented code completion framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6227–6240, Dublin, Ireland. Association for Computational Linguistics.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. MoleculeQA: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3769–3789, Miami, Florida, USA. Association for Computational Linguistics.

Man Luo, Shashank Jain, Anchit Gupta, Arash Einolghozati, Barlas Oguz, Debojeet Chatterjee, Xilun Chen, Chitta Baral, and Peyman Heidari. 2023. A study on the efficiency and generalization of light hybrid retrievers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1617–1626, Toronto, Canada. Association for Computational Linguistics.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2024. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*.

Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. 2025. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning. *arXiv preprint arXiv:2502.00848*.

Daphne E. Smith Marsh. 2023. Overview of generic drugs and drug naming.

Nashwa Masnoon, Sepehr Shakib, Lisa Kalisch-Ellett, and Gillian E Caughey. 2017. What is polypharmacy? a systematic review of definitions. *BMC geriatrics*, 17:1–10.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 2025. Laypeople's use of and attitudes toward large language models and search engines for health queries: Survey study. *Journal of Medical Internet Research*, 27:e64290.

Meta. 2024. Llama 3.3 | Model Cards and Prompt formats — llama.com. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. [Accessed 01-05-2025].

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, and 1 others. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, Jungwei Wilfred Fan, and 1 others. 2023. Extractive clinical question-answering with multianswer and multifocus questions: data set development and evaluation study. *JMIR AI*, 2(1):e41818.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and leman go neural: higher-order graph neural networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Neo4j. 2025. Neo4j Graph Database & Analytics – The Leader in Graph Databases — neo4j.com. https://neo4j.com/. [Accessed 13-05-2025].

National Library of Medicine. 2025. Home - Gene - NCBI — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/gene. [Accessed 09-05-2025].

OpenAI. 2024. OpenAI Platform — platform.openai.com. https://platform.openai.com/docs/models/text-embedding-3-large. [Accessed 13-05-2025].

OpenAI. 2025a. Introducing GPT-4.1 in the API — openai.com. https://openai.com/index/gpt-4-1/. [Accessed 11-05-2025].

OpenAI. 2025b. OpenAI o3 and o4-mini System Card — openai.com. https://openai.com/index/o3-o4-mini-system-card/. [Accessed 01-05-2025].

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Junwoo Park, Youngwoo Cho, Haneol Lee, Jaegul Choo, and Edward Choi. 2021. Knowledge graph-based question answering with electronic health records. In *Machine Learning for Healthcare Conference*, pages 36–53. PMLR.

Helen Pearson. 2006. Genetics: what is a gene? *Nature*, 441(7092).

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *NeurIPS*.

LibreTexts project. 2020. Line Notation (SMILES and InChI). [Online; accessed 2025-04-23].

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shaina Raza, Brian Schwartz, and Laura C Rosella. 2022. Coquad: a covid-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC bioinformatics*, 23(1):210.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

sentence transformers. 2021. sentence-transformers/all-MiniLM-L6-v2 · Hugging Face — huggingface.co. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Accessed 13-05-2025].

Rotem Shalev-Arkushin, Rinon Gal, Amit H Bermano, and Ohad Fried. 2025. Imagerag: Dynamic image retrieval for reference-guided image generation. *arXiv preprint arXiv:2502.09411*.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Kate Silver. 2020. Ever Wonder How Drugs Are Named? Read On | Pfizer — pfizer.com. https://www.pfizer.com/news/articles/ever_wonder_how_drugs_are_named_read_on. [Accessed 22-04-2025].

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023(5):1–95.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Duxin Sun. 2022. 90 https://www.asbmb.org/asbmb-today/opinions/031222/90-of-drugs-fail-clinical-trials. [Accessed 13-05-2025].

Yalin Sun, Shu Teng, Wen Li, Huaping Wang, and Zhenghong Qi. 2025. Genetic analysis of a female patient with x-linked lymphoproliferative disease type 2: a case report. *Journal of Medical Case Reports*, 19(1):193.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, and 1 others. 2023. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. 2024. MolLM: a unified language model for integrating biomedical text with 2D and 3D molecular representations. *Bioinformatics*, 40(Supplement_1):i357–i368.

Qwen Team. 2025. Qwen3.

Thiviyan Thanapalasingam, Lucas van Berkel, Peter Bloem, and Paul Groth. 2022. Relational graph convolutional networks: a closer look. *PeerJ Computer Science*, 8:e1073.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

USMLE. 2025. Home | United States Medical Licensing Examination — usmle.org. https://www.usmle.org/. [Accessed 13-05-2025].

Sriram Vajapeyam. 2014. Understanding shannon's entropy metric for information. *Preprint*, arXiv:1405.2061.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Caterina Vicidomini and Giovanni N Roviello. 2023. Protein-targeting drug discovery.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025a. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. 2025b. Retrieval-augmented perception: High-resolution image perception meets visual rag. *arXiv preprint arXiv:2503.01222*.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024b. Beyond the limits: a survey of techniques to extend the context length in large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. 2023. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and RISHITA ANUBHAI. 2024c. Biobridge: Bridging biomedical foundation models via knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Wikipedia contributors. 2025. Pubmed — Wikipedia, the free encyclopedia. [Online; accessed 13-May-2025].

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. 2024a. Synthetic multimodal question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA. Association for Computational Linguistics.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis Ioannidis, Karthik Subbian, James Y Zou, and Jure Leskovec. 2024b. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37:127129–127153.

Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. 2024. Bridging the gap: Generating a comprehensive biomedical knowledge graph question answering dataset. In *ECAI 2024*, pages 1198–1205. IOS Press.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312, Singapore. Association for Computational Linguistics.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, and 1 others. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Dharunish Yugeswardeenoo, Kevin Zhu, and Sean O'Brien. 2024. Question-analysis prompting improves LLM performance in reasoning tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 402–413, Bangkok, Thailand. Association for Computational Linguistics.

Hye Sun Yun and Timothy Bickmore. 2025. Online health information–seeking in the era of large language models: Cross-sectional web-based survey study. *Journal of Medical Internet Research*, 27:e68560.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Renjun Xu, Hongyang Chen, Xiaohui Fan, and 2 others. 2025. Scientific large language models: A survey on biological & chemical domains. *ACM Comput. Surv.*, 57(6).

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024a. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and 1 others. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.

A Knowledge Base Sampling

1. The base knowledge graph (c.f. §3.1.1) contains ~4M drug-drug interactions. However, as we explain in Appendix D.1, most drugs do not have common/generic names. We require our drug entities to have common names for acquiring background-text data to generate questions (c.f. §3.4). As such, we sample a set of DDIs yielding a total of 494 drugs with ~18.5K edges.

2. The total number of proteins interacting with each of the 494 drugs is 198, yielding ~15K DPI edges. However, most of these edges are absent in the STITCH database (Appendix C). This ultimately leaves a total of 314 DPI edges.

3. We explain why there are no protein-protein (PPI) edges in Appendix H.

Our resulting graph thus contains ~18.5K DDIs and 314 DPIs, yielding a total of ~18.9 edges, 494 drug and 198 protein nodes.

B SMILES

When querying PubChem for drug data, it returns two kinds of `SMILES`, *canonical* and *isomeric*. Various possibilities exist for a compound's potential `SMILES`. The canonical `SMILES` represents a unique identifier for a given compound, while the isomeric `SMILES` includes additional stereochemistry and isotope information. We decided to use only the canonical `SMILES` as we are more interested in the bare-bones molecular structure over more nuanced information as provided by the isomeric `SMILES`.

C STITCH Database

We use the subset of `STITCH` on drug-protein interactions related to humans, which can be downloaded from http://stitch.embl.de/download/actions.v5.0/9606.actions.v5.0.tsv.gz. All interactions are associated with a *score* (numeric value indicating the strength of the interaction), *mode* (activation, phenotype, binding, predicted to bind, catalysis, inhibition, reaction) and *action* (activation or inhibition). In simple terms, *mode* means the type or nature of interaction between the drug and protein, while *action* refers to the outcome of the interaction[1]. It should be noted that *action* is *directional* in nature, i.e., to show which entity acts on which. This adds a directional component to our graph.

For each DPI in our graph (15K), we query the `STITCH` database to acquire the edge labels. If the DPI does not exist in STITCH, we discard it. If it exists, STITCH returns a (score, mode, action) triple for the DPI. We retain only those edges whose score is greater than 900, i.e., indicating an extremely likely or strong interaction. This ultimately results in 314 DPIs.

As mentioned, a DPI has a *mode* and *action*. To obtain a unified label, we combine the two fields. This results in our final DPI label set of *binding and inhibition* and *binding and activation*, which indicates that a drug *binds* to a protein and either *inhibits* or *activates* its (proteins) function.

D Resolving Entity Names

D.1 Resolving Drug Names

The process of naming a drug is a long and arduous one. After discovery, a potential drug compound is first registered in a medical database, such as PubChem [2], and then goes through several layers of clinical trials to test for efficacy and ultimately marketability where it gets a generic name like, *Fluoxetine* (Marsh, 2023; Silver, 2020). Unfortunately, most drugs do not make it through such rigorous testing, which can often take as long as 10 years (Marsh, 2023). Thus, despite being registered, we do not find much information on most compounds.

For our purposes, we require a drug to have been studied and documented for utilisation in our dataset. As such, we follow a multi-step process to resolve drug names as shown in Figure A1 to get the most coverage. A drug is represented in our database by a CID or PubChem Compound Identification such as `CID000002088` (*alendronic acid*). First, we query its entry in the PubChem database to obtain its International Chemical Identifier (InChI) key, a unique compound identifier shared across databases. Using this key, we query the ChemSpider database [3] as its API has a feature to access a drug's common name. If this fails, we return to PubChem and determine if a set of drug *synonyms* (names used across literature or clinical trials) exists.

If no synonym exists, we consider the IUPAC (International Union of Pure and Applied Chemistry) name (a long standardised name describing

[1] http://stitch.embl.de/download/README

[2] https://pubchem.ncbi.nlm.nih.gov/

[3] https://www.chemspider.com/

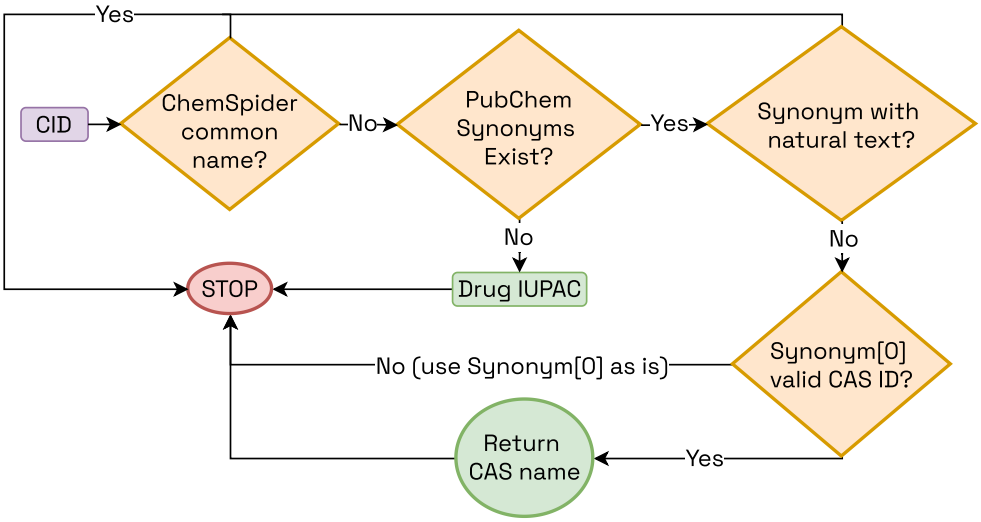

Figure A1: Flowchart for resolving drug names.

the entire molecular structure) as the drug's name. Else, we try to find any synonym with *natural text*, i.e., no non-alphabet characters (numbers, hyphens, etc.). If there exists no such synonym, then we consider the first synonym in the returned list as the common name, as we observe that, more often than not, the first entry is a reasonably well-structured string. Also, we find that the first synonym may also be a CAS (Chemical Abstracts Service) [4] number, which gives us a final chance to obtain a common name for the drug.

D.2 Resolving Protein Names

Acquiring protein names is more straightforward than acquiring drug names. In the base knowledge-graph (c.f. §3.1.1), proteins are represented by their source *gene*, molecules that encode information for synthesising proteins (Pearson, 2006). A gene is provided as an NCBI (National Center for Biotechnology Information) Entrez database entry[5] such as `3351` (`HTR1B`, the gene name corresponding to the given ID). As our goal is to acquire the protein associated with these genes, we query STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)[6], a database tailored for protein interactions. A useful feature of the STRING API is that, given a gene ID, it automatically maps it to the correct or main protein in its database. This helps us resolve each gene to its corresponding protein name. For example, `HTR1B` (gene) gets mapped to `5-hydroxytryptamine receptor 1B` (protein).

[4] https://www.cas.org/cas-data/cas-registry
[5] https://www.ncbi.nlm.nih.gov/gene
[6] https://string-db.org/

E Additional Related Work

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) aims to enhance the accuracy of LLMs by grounding their responses to appropriate supporting knowledge relevant to a given query. The most popular knowledge source used in current RAG methods is unstructured text documents (Fan et al., 2024; Gao et al., 2023). Each method then differs in either what document is retrieved or how the LLM utilises the retrieved document. For example, IRCoT (Trivedi et al., 2023) performs a sequence of retrieval + reasoning steps based on intermediate LLM responses; Self-RAG (Asai et al., 2023) forces the LLM to critique its response based on the retrieved documents, etc. It should be noted that given the diversity in RAG frameworks, there is no one outright state-of-the-art (SOTA) (Jin et al., 2024). This is because it depends on the target task, where using advanced retrieval methods such as FLARE (Jiang et al., 2023b) can be outperformed (Jin et al., 2024) by naive RAG (Lewis et al., 2020) as it forces the LLM or retriever to overanalyse a problem.

Switching gears from text, we do find frameworks that have begun incorporating different data modalities for solving tasks such as image captioning (generating text descriptions of images (Stefanini et al., 2022)) and code completion (Lu et al., 2022). However, these frameworks typically target retrieval and reasoning over a single modality (Mei et al., 2025; Jin et al., 2024; Abootorabi et al., 2025) such as text (Trivedi et al., 2023; Asai et al., 2023; Shi et al., 2024; Izacard et al., 2023), knowledge

graph (Edge et al., 2024; He et al., 2024; Mavromatis and Karypis, 2024; Hu et al., 2024a), or image (Shalev-Arkushin et al., 2025; Lyu et al., 2025; Wang et al., 2025b; Chen et al., 2022a).

Over the next five sections, we describe studies relevant to our two overarching themes, viz., RAG [(i) studies that have applied RAG to medical applications (ii) datasets to support RAG testing] and QA [(iii) QA datasets over molecular data (iv) modalities seen in medical-QA datasets and (v) multi-modal QA datasets]. Each of these topics connects to a specific aspect of our dataset.

RAG In Medicine Although most RAG pipelines are tested on general domain data (Jin et al., 2024; Abootorabi et al., 2025), there are studies applying RAG for biomedical tasks. Lozano et al. (2023) builds an application for QA over PubMed (Wikipedia contributors, 2025) articles. Wang et al. (2023) performs molecule synthesis by retrieving similar samples from a labelled database. However, they follow the *unimodal* setup, i.e., limiting retrieval to a single source of information. Although Wang et al. (2024c) performs QA over a multi-modal knowledge graph of biological entities, including drugs and proteins, the questions themselves do not integrate each source, essentially collapsing into unimodal retrieval. (Zhu et al., 2024) investigates patient readmission/mortality by utilising both time-series and text data. However, they retrieve auxiliary text relative to each modality, thus resulting in a text-only retrieval framework.

RAG Datasets Current benchmarks for RAG, such as RAGBench (Friel et al., 2024) and ChatRAG Bench (Liu et al., 2024c), are a combination of existing QA datasets. The original QA datasets were transformed to fit RAG testing by augmenting them with a corpus. That said, we do find two new datasets that directly support RAG, viz., CRAG (Yang et al., 2024) and STaRK (Wu et al., 2024b). CRAG only discusses unimodal questions related to the general domain (sports, music, etc.). The dataset closest to our work is `STaRK` (Wu et al., 2024b). Here, questions are designed by combining knowledge graphs with text data for the nodes. We find two limitations with their dataset. First, a "relation" is encountered only when querying the graph with a *template* without propagating it to the final question. Second, even though the answers are nodes in the graph, only the gold node's document is used to write questions, while the other nodes' texts are ignored. This precludes the LLM from accessing valuable information associated with the other node.

Molecular QA Recently, QA based on molecules has seen growth. In this regard, we find two datasets, `MoleculeQA` (Lu et al., 2024) and `PubChemQA` (Luo et al., 2024). Each dataset has limitations. Questions in both are template-based, i.e., they have fixed patterns. `PubChemQA` asks *describe this molecule* while `MoleculeQA` has types such as *Which kind of compound does this molecule belong to?*. This heavily limits sample diversity. Furthermore, neither dataset considers relationships between molecules, and they restrict themselves to one modality by ignoring background knowledge of the molecules.

Modalities For Medical-QA In medical QA, we typically see datasets dealing with one of three modalities exclusively (Jin et al., 2022), i.e., *text* (Moon et al., 2023; Raza et al., 2022), *knowledge graphs* (Park et al., 2021; Yan et al., 2024) or *images* (Zhang et al., 2024b; Hu et al., 2024b). However, as mentioned before, medical data covers a wide range of modalities. Limiting attention to a single source, as these datasets, limits the generalisation of the developed models.

Multi-Modal QA Apart from STaRK, we find related datasets for multi-modal QA/RAG including `ProMQA` (Hasegawa et al., 2024), `MultiModalQA` (Talmor et al., 2021) and `SPIQA` (Pramanick et al., 2024). The first two datasets either create questions combining multiple modalities (Talmor et al., 2021) or involve them for reasoning tasks (Hasegawa et al., 2024). However, they focus on the general or common-sense domain, where recent studies (Wei et al., 2022; Krause and Stolzenburg, 2024) have shown LLMs to be proficient in. This leads to a testbed not complex enough for evaluation. The third dataset asks questions about scientific papers (images) and tables. Although they use two modalities, their questions are either-or, i.e., either about images or tables, without connecting the two.

`BioMol-MQA` stands at the intersection of these areas. It realises the necessity for retrieving multi-modal healthcare information across diverse modalities, as well as questions that target an LLM's abilities to reason across domains. By targeting the limitations of the above studies, we aim to provide a valuable resource for complex reasoning and retrieval in the form of `BioMol-MQA`.

Prompt For Text Post-Processing

You are an expert in drug biology. Given source material on a specific drug, please rewrite it in one paragraph suitable for an expert audience.
TASK GUIDELINES:

1. Do your best to make the material as dense as possible by using complex domain-specific jargon.
2. DO NOT ADD any external information, i.e., strictly utilise the provided context only.
3. Ignore historical details in the text, such as year of discovery, etc., focusing only on the core biological details.
4. Avoid including unnecessary text such as "Here are the key points" or other filler language.

DRUG NAME: {}
DRUG INFORMATION: {}

Figure A2: Text Post-Processing Prompt

Prompt For Molecular Interaction Extraction

You are an expert in molecular chemistry and pharmacology. Given the molecular structures of two drugs, represented by their SMILES strings, identify one specific molecular interaction between them using only the structural information provided.
TASK GUIDELINES:

1. Focus only on interactions relevant under physiological conditions (e.g., hydrogen bonding, steric clashes, electrostatic interactions).
2. Do not infer interactions from external knowledge or assumptions about the drug identities.
3. Only report interactions supported by structural features in the SMILES.
4. If no interaction exists, respond with 'NONE'.

For any interaction you do identify, format your response as:

- INTERACTION: [Specific name of the interaction]
- MECHANISM: [Brief explanation of how/why this interaction occurs]
- EVIDENCE: [Direct structural features or groups from the SMILES that support this]
- SEVERITY: [Low / Moderate / High - based on likely pharmacological impact]

SMILES 1: {}
SMILES 2: {}

Figure A3: Molecular Interaction Extraction Prompt

Prompt/Rubric For Question Evaluation

You are a judge who evaluates Question-Answer (QA) pairs based on drug-drug interactions. The following information was used to create the questions,

- Background text on each drug.
- A knowledge-graph triple (in subject-predicate-object format) that explains the side effect of taking two drugs together.

A question incorporates knowledge-graph triples and background text. Given the necessary information, please evaluate the QA-pair using the following rubric,

1. **CLARITY**: How difficult is the question's language?

   SCORES:

   (a) 0 - Easy; Straightforward and uses common phrases like "muscle-pain".
   (b) 1 - Medium; Overall comprehensible but uses some domain-specific jargon like "hepatotoxicity".
   (c) 2 - Hard; Quite difficult and requires good domain expertise to answer.

2. **COVERAGE**: Does the question make use of the provided modalities (text/triple)? Note, the question does *NOT* need to utilise *ALL* of the given text.

   SCORES:

   (a) 0 - Low; Completely ignores the given information.
   (b) 1 - Medium; Uses only information for one entity or one modality.
   (c) 2 - High; Each entity and modality receives decent coverage.

3. **ASSUMPTIONS**: Does the question include information beyond what is provided?

   SCORES:

   (a) 0 - Bad; A lot of the question's data is absent from the provided data.
   (b) 1 - Okay; Most of the question relies on the provided data, but some assumptions are made.
   (c) 2 - Best; Relies strictly on the provided data.

4. **INFERABLE**: Can the answer be derived from the provided information?

   SCORES:

   (a) 0 - No; The answer is irrelevant in regard to the question.
   (b) 1 - Maybe; Can be potentially derived but requires additional data to infer.
   (c) 2 - Yes; There exists entailment between the answer and the question and its associated context.

TASK REQUIREMENTS:

1. Do not write filler text such as "Here is my evaluation", etc.
2. Provide your output as,

   - METRIC REASONING: <Brief explanation of thought process for the metric.>
   - METRIC SCORE: <0, 1, 2 based on the guidelines. No need to repeat the descriptions for each score.>"""

BACKGROUND INFORMATION:{}
QUESTION: {}
ANSWER: {}

Figure A4: Prompt for LLM-based evaluation. Our human evaluator was provided with just the rubric and not the prompt.

Prompt For 1-hop DDI (Bio) Question Generation

You are a helpful AI assistant tasked with generating one question about a drug-drug interaction (DDI) based on background information about two drugs and a knowledge-graph triple describing their interaction (subject-predicate-object format).
TASK REQUIREMENTS:

1. Write exactly one question integrating the background knowledge of both drugs and their relationship. The answer must be either Drug 1 or Drug 2.
2. The question may be as complex as desired, but it must be answerable.
3. Do NOT mention the drugs by name in the question; use only their background descriptions.
4. The question should **specifically test knowledge of the triple-described relationship or interaction, not just isolated facts about either drug.**
5. The answer should be only the name of the correct drug.
6. Output in the following format:

   Question:
   Answer:

DRUG 1 NAME: {}
DRUG 1 BACKGROUND INFORMATION: {}

DRUG 2 NAME: {}
DRUG 2 BACKGROUND INFORMATION: {}

DRUG-DRUG INTERACTION TRIPLE (subject-predicate-object): {}

Figure A5: Question Generation Prompt For Bio-based 1-hop DDIs.

F Prompts Used

An overview of each prompt is given in Figures A2, A3, A4 and A5.

G Reasons For Not Using IRCoT/Gemini

We utilise standard RAG (Lewis et al., 2020) for our experiments. Although newer variants exist, including IRCoT (Interleaving Retrieval with Chain-of-Thought), we avoid them for the following reasons:

1. Standard RAG still shows promise against IRCoT (Jin et al., 2024) often outperforming it and other variants on different datasets.

2. IRCoT is a sequential framework, meaning that each step of reasoning depends on the prior retrieval and reasoning step. This increases the overall number of API calls, making it expensive to implement. For context, all of our generations were done in batch processing mode to limit costs. Synchronously calling API's incurs more overhead.

3. As a result of the sequential nature of IRCoT, batch processing is not possible, thereby limiting scalability.

Finally, we do attempt to use Google's `Gemini 2.0 Flash` (Google, 2024). However, the results were quite poor, and we noticed that it struggled to follow our prompts. As such, we did not use it in our experiments.

H Absence Of Protein-Protein Interactions

There are several protein-protein interactions in the base knowledge graph (Zitnik et al., 2018). However, like DPIs, they are unlabeled and require us to query STRING to get the associated data. When two genes, such as `114787` and `2775`, are queried, STRING returns a dictionary of their associated proteins and interactions as follows,

```
{
    "stringId_A": "9606.ENSP00000262494",
    "stringId_B": "9606.ENSP00000305839",
    "preferredName_A": "GNAO1",
    "preferredName_B": "GPRIN1",
    "ncbiTaxonId": 9606,
    "score": 0.635,
    "nscore": 0,
    "fscore": 0,
    "pscore": 0,
    "ascore": 0.083,
    "escore": 0,
    "dscore": 0,
    "tscore": 0.618
}
```

Each score is a source of evidence for the protein-protein interaction (PPI). For example, `tscore` (text score) means that a relationship exists between the two proteins as documented by existing studies, etc. If we want to convert these scores to a natural language label, it raises questions such as, which score to choose, should we choose multiple scores or the aggregate, etc. As there is no natural way to acquire a label from these scores, we decide not to include PPIs in our dataset.

I Question Examples

Figure A6 shows examples of questions based on molecular-level drug-drug interactions as well as drug-protein interactions.

J Additional RAG Results

We provide results for the other LLMs using our multi-modal retriever in Table A1. As the scores indicate, each model benefits from having the associated domain knowledge. Even weaker models like Mistral show enhanced capabilities when provided with the necessary information. Although the gap between the smaller LLMs and frontier models such as Claude/DeepSeek is big, the trend with all of them is constant, i.e., having the necessary background data (even from a simple multi-modal retrieval) helps them to reason over multi-modal contexts.

K Dense Text Retrievers

We consider four dense retrieval models, all using a BERT-style backbone, except OpenAI's `text-embedding-3-large` for which we do not have training details.

- `MedCPT` is specifically tailored to the medical domain by extensively tuning a `PubMedBERT` (Gu et al., 2021) checkpoint for query-article retrieval.

- `MolLM` is an LLM-based retriever that fine-tunes a BERT checkpoint on text, `SMILES`, and molecular graphs. As such, its inclusion provides insights into retrieval for `SMILES` data.

DRUG NAME: Loxapine
DRUG BACKGROUND INFORMATION: ... is a tricyclic antipsychotic medication used primarily in the treatment of schizophrenia ... and structurally very similar to clozapine ...

PROTEIN NAME: Muscarinic acetylcholine receptor M4
PROTEIN BACKGROUND INFORMATION: ... cholinergic receptor ... is a protein that, in humans, is encoded by the CHRM4 gene.

DRUG-PROTEIN INTERACTION: Loxapine-binding and inhibition-Muscarinic acetylcholine receptor M4

Q: Which protein, encoded by the CHRM4 gene and functioning as a cholinergic receptor subtype, is bound and inhibited by a tricyclic antipsychotic medication primarily used for schizophrenia that is structurally similar to clozapine?
A: Muscarinic acetylcholine receptor M4

DRUG 1: Alprazolam
DRUG 1 SMILES: CC1=NN=C2N1C3=C(C=C(C=C3)Cl)C(=NC2)C4=CC=CC=C4

DRUG 2: Ciclopirox
DRUG 2 SMILES: CC1=CC(=O)N(C(=C1)C2CCCCC2)O

DRUG-DRUG INTERACTION: Alprazolam-Hydrogen Bonding-Ciclopirox

---

Q: Which drug, the one containing a triazolobenzodiazepine core fused to a chlorinated aromatic ring, acts as the hydrogen bond donor or acceptor in its interaction with the molecule featuring a cyclohexylpyridone structure due to potential hydrogen-bonding functional groups?
A:Ciclopirox

Figure A6: Question based on DPI (Top) and DDI (molecular) (Bottom).

Model	Closed Source LLMs		Open Source LLMs				
Approach	o4-mini	Claude 3.7 Sonnet	DeepSeek R1	LLama 3.3	Qwen 3	TxGemma	Mistral
Zero-Shot	(0.32, 0.42, 0.79)	(0.30, 0.36, 0.78)	(0.33, 0.40, 0.87)	(0.27, 0.34, 0.85)	(0.22, 0.25, 0.51)	(0.07, 0.16, 0.82)	(0.02, 0.14, 0.81)
RAG	**(0.54, 0.62, 0.92)**	(0.46, 0.46, 0.87)	**(0.48, 0.53, 0.88)**	(0.38, 0.42, 0.86)	(0.41, 0.44, 0.57)	(0.32, 0.38, 0.84)	(0.20, 0.34, 0.84)
Upper Bound	(0.88, 0.88, 0.94)	(0.89, 0.89, 0.94)	(0.90, 0.90, 0.98)	(0.71, 0.71, 0.93)	(0.80, 0.80, 0.88)	(0.76, 0.76, 0.94)	(0.74, 0.76, 0.95)

Table A1: Additional RAG Results. **Bold** represents the best performing model in each category for the RAG test. Each tuple is (lexical EM, lexical F1, BERTScore F1)

- `DPR` is a simple BERT-based retriever trained for question answering tasks.
- `text-embedding-3-large` (OpenAI, 2024) is our only proprietary model. As such, we do not have details on how it was trained and can only use its reported results as (OpenAI, 2024) evidence for its utility.

L Graph Retrievers

As we deal with graphs, it makes sense to test graph retrievers. Training graph neural network (GNN) based retrievers is not as straightforward as text-based retrievers due to the ingrained differences in structure (edges, features, etc.) across graphs (Hou et al., 2024). That said, we attempt to train three GNN-based retrievers, i.e., Graph Convolutional Network (GCN) (Morris et al., 2019), Graph SAGE (Hamilton et al., 2017), and Graph Attention Network (Veličković et al., 2018) for our task. To train the models, we utilise the entire graph (c.f. Table 1) along with the questions in the training split of our dataset and evaluate on the validation split. The idea here is to train a GNN to retrieve triples (entity-relation-entity) by jointly optimising triple and query embeddings. Unfortunately, due to the size of our training data and graph, we are unable to learn useful representations, as GNNs are usually trained on much larger samples (Leskovec and Krevl, 2014). Table A2 gives the results of our graph retrievers.

Due to the size of the graph, we treat it as a homogeneous structure (single entity/relationship type), similar to the base graph learning method from (Zitnik et al., 2018). Treating our knowledge graph as a heterogeneous graph (multiple node/edge types) and using RGCN (Relational-GCN) layers (Schlichtkrull et al., 2018) did not seem to work either. This makes sense as RGCN networks typically work well with larger data and a higher number of edges per relationship (Thanapalasingam et al., 2022). We try to *hot start* our optimisation, by pretraining a GNN for a *link prediction* (Li et al., 2023) (detecting if an edge exists between nodes). However, this also yielded similar results as above.

Finally, we tuned various hyperparameters and network specifications, which also did not help. Ultimately, we settled on each GNN being a two-layer network, with ReLU activation. They were trained to minimise triplet loss, i.e., given a triple of (query, positive triple, negative triple), a model learns to minimise the distance between the query and positive triple (from our ground truth set) while maximising the distance between the query and negative triple. We train our models for 10 epochs with an AdamW optimiser (Loshchilov and Hutter, 2019) and a learning rate of $1e-4$. We use `all-MiniLM-L6-v2` (sentence transformers, 2021) as our query encoder.

L.1 Neo4j retriever

We provide a more detailed explanation of how Neo4j operates, as follows,

1. It first builds indexes, i.e., builds a graph database using the provided set of triples (entity-relation-entity).
2. Next, it creates embeddings (using `all-MiniLM-L6-v2` (sentence transformers, 2021)) for each relation in the database to locate relationships semantically similar to the one described by the question.

When a query arrives,

1. It first does a best attempt to find entities in the question via Named Entity Recognition (implemented with Flair models (Akbik et al., 2019))
2. Next, it looks for relations closest to the query using the same sentence encoder (`all-MiniLM-L6-v2` (sentence transformers, 2021)) used to embed the relations.
3. Finally, it returns the top-k triples that have these entities and relationships.

Model	Hits@5	Hits@10	Hits@15	Recall@5	MRR
Neo4j	0.04/0.04	0.07/0.08	0.07/0.09	0.01	0.03
Graph Convolution Network	0.01/0.02	0.02/0.04	0.03/0.06	0.01	0.01
Graph SAGE Network	0.01/0.02	0.01/0.04	0.01/0.06	0	0.01
Graph Attention Network	0.02/0.05	0.04/0.08	0.04/0.09	0.01	0.02

Table A2: Neo4j + GNN retrievers trained on the entire graph. Scores are based on the test set. Hits are reported as hard/soft hits.

All of these steps can be modified. We tried different settings (such as using a simpler regex-based entity recogniser, different embedding models such as PubMedBERT (Gu et al., 2021), etc.) and landed on this version, which provided decent results.

M Question Analysis

We use four quantitative metrics, averaged over all questions, for analysis (c.f. Table 4) as described below,

1. **Question Length**: The average question length (in tokens[7]) is much higher than existing biomedical QA datasets such as PubMedQA (Jin et al., 2019) and even datasets designed specifically for long-context QA such as QuALITY (Pang et al., 2022) (66 v/s 14 and 12.5 respectively) indicating a need for deep reasoning and tracking of facts across multiple segments for our questions. The overall distribution of question lengths is shown in Figure A7. Considering the complex and detailed nature of our queries, especially multi-hop ones, such question lengths are justified as being almost akin to medical case reports (Sun et al., 2025).

2. **Type-to-Token (TTR) Ratio**: TTR (Shen, 2022) is defined as the (total number of unique words/total words) in a text piece. It is used to gauge linguistic diversity. Lower values imply repetitive text, i.e., a small number of unique words and vice versa. With an average TTR of 0.84, our questions display a strong spread of vocabulary, beneficial for training models to learn technical/jargon-dominated language in domains such as ours.

3. **Shannon Entropy**: We measure Shannon Entropy (Vajapeyam, 2014) to gauge semantic coverage. In other words, this measure will tell us if the questions are repetitive or diverse in content. Entropy (in bits) lies between 0 and $\log_2(k)$ where k is the number of categories. In our case, we use question length (tokens) as categories, which gives 6.69 as the maximum entropy. An average of 5.67 thus indicates a high semantic spread in our questions.

4. **Dependency Tree Depth**: A dependency tree[8] (Jurafsky and Martin, 2025) provides an overview of the grammatical roles (determiner, noun-phrase, etc.) of each word and the relationships (subject, object, etc.) that exist between them. We consider the average dependency tree depth for all questions (10.86) as a measure of linguistic difficulty. Longer sentences have deeper trees due to more grammatical structures. As our average question length is ~66 tokens, such depths make sense. This highlights the linguistic challenges posed by our questions.

As for question *types*, we find only three categories, *Which* (1650 questions), *What* (30) and *Identify* (3). It should be noted that this distribution is not based on the first word of the question, but rather where the interrogative phrase occurs. Analysing questions based on the first word reveals more variety, such as questions starting with *A*, *Between*, *Considering*, etc. While these types are similar to factoid-style questions (Rajpurkar et al., 2016), their content and source distinguish them from other QA datasets.

Qualitative analysis by our human evaluator reveals that our *1-hop questions are harder* to answer than the 2/3-hop ones. This counterintuitive observation is explained by the fact that the latter category of questions has higher information density, or more content to learn from. Overall, their

[7]All token measurements used throughout the paper are done using the GPT-4o tokenizer.

[8]Implemented using spaCy (`https://spacy.io/`).

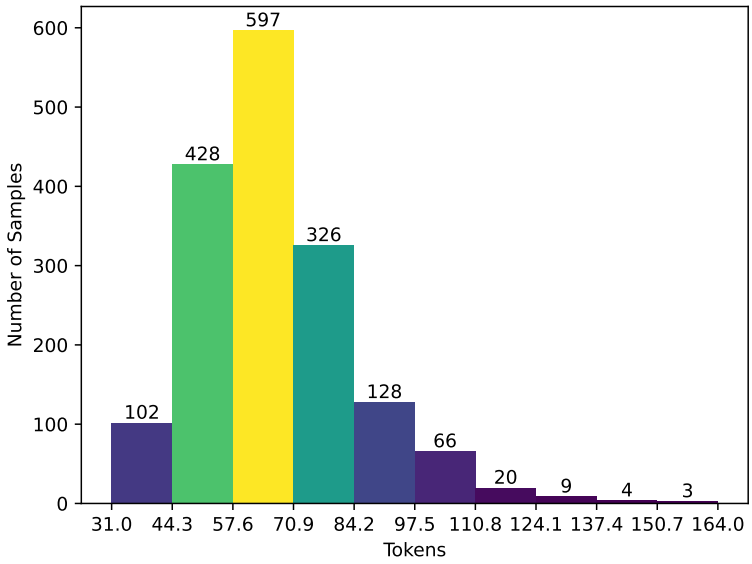

Figure A7: Question Length Distribution based on tokens.

assessment rated the questions as fair and even suitable for advanced (5th year) students preparing for the USMLE (USMLE, 2025) (exam to acquire one's medical license).

N LLM Settings

None of our LLMs were trained, i.e., we only show results on the test set of our dataset through LLM inference. We have a total of 7 LLMs (o4-mini, Claude 3.7 Sonnet, DeepSeek R1, LLama 3.3(Llama-3.1-8B-Instruct-Turbo), Qwen 3(Qwen/Qwen3-235B-A22B-fp8-tput), TxGemma(TxGemma-9B-predict), Mistral (Mistral-7B-Instruct-v0.2). Except for TxGemma, each model is run using default settings as provided by the respective APIs, i.e., we do not make any modifications to the temperature, top_k, etc. TxGemma is run on our hardware as it is not available via API. However, even in this case, we use it out-of-the-box without changing any settings. To run our dense retrievers as well as TxGemma, we utilise a single NVIDIA A6000 48GB card.

O Code and Data

Our code and dataset will be released upon publication of this paper.