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

Saptarshi Sengupta, Shuhua Yang, Paul Kwong Yu, Fali Wang, Suhang Wang
Pennsylvania State University
{sks6765,sky5341,pky5070,fqw5095,szw494}@psu.edu
Code 🤗 Dataset

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 [92] have shown great performance in various tasks [69], such as text summarization [91] and question answering [68]. Their success has attracted more and more people to adopt them for daily use, e.g., using LLM-based chatbots to address medical concerns [54, 88]. Despite their strong abilities, LLMs also face issues such as hallucination [28], knowledge cutoff [12] and lacking domain-specific knowledge [85, 19]. Those issues hinder the adoption of LLMs in high-stakes scenarios such as healthcare and finance, where incorrect responses could mislead end users and put their finances, health, and lives at risk.

One efficient and popular way to address these issues is Retrieval Augmented Generation (RAG) [40], which retrieves context relevant to a query and has an LLM ground its response on retrieved information for factual accuracy. Various RAG methods [15, 17] have emerged over time. However, they mainly focus on retrieving information from a single modality [32], e.g., text [74], knowledge graph [14], or image [67] only. Recently, there have been some initial attempts for multi-modal RAG [3], i.e., incorporating more than one modality for retrieval. However, they mainly focus on general knowledge domains [32, 3] that do not require expertise in any particular area. For many real-world problems, e.g., medical and healthcare, information relevant to medical queries can manifest in various modalities such as knowledge graph, text (clinical notes), and complex molecular structure from crystallography. As such, being able to retrieve the relevant domain-specific information from

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

multi-modality, and reason and synthesize diverse knowledge to generate an accurate response is crucial for LLM. However, the work on understanding and developing LLMs with such ability is rather limited [6]. 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 question answering on polypharmacy. Polypharmacy [24, 53], 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, if drug-drug interactions are not known, it may exacerbate rather than improve one's health. Given the seriousness of the phenomenon and the extensive domain knowledge required for understanding and answering questions related to polypharmacy, polypharmacy is a perfect testbed to gauge an LLM's multi-modal reasoning abilities. Here, an LLM must have access to 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 [75, 8]. Having access to this modality (protein-level data) will help in better resolving polypharmacy queries.

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, pinklorange = Text) that a model must connect to answer questions. In this example, a model 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 are in App. 9.

`BioMol-MQA` is composed of two parts: (i) a multimodal knowledge graph (KG) for information retrieval; and (ii) complex medical questions that require multimodal RAG and LLM reasoning for question answering. The multimodal KG provides complex and comprehensive domain knowledge in three modalities: a knowledge graph 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 these questions, an LLM must retrieve relevant information from the three modalities to provide an accurate response. To facilitate evaluation and model fine-tuning, each question has groundtruth answer and resources from which the problem was constructed (§3.4). An example question is shown in Figure 1. The task in our benchmark is formally defined 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) [43] or molecular property prediction [23].

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 [32, 3], there are studies applying RAG for biomedical tasks. Lozano et al. [49] builds an application for QA over PubMed [82] articles. Wang et al. [79] perform molecule synthesis by retrieving similar samples from a labeled database. However, they follow the *unimodal* setup, i.e., limiting retrieval to a single-source of information. Although Wang et al. [80] 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.

RAG Datasets Current benchmarks for RAG, such as RAGBench [16] and ChatRAG Bench [47], 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 [86] and STaRK [84]. CRAG only discusses unimodal questions related to the general domain (sports, music, etc.). STaRK combines questions from graph and text modality to support retrieval over a multi-modal graph. However, the issue with STaRK is in its question construction. For each KG triple (entity-relation-entity), they only consider the text-background of one entity and do not *strongly* integrate relationships into their questions (App. 5). This limits the complexity and coverage of their questions.

Molecular QA Recently, QA based on molecules has seen growth. In this regard, we find two datasets, `MoleculeQA` [50] and `PubChemQA` [52]. 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.

`BioMol-MQA` addresses the above limitations by effectively integrating multiple modalities, considering relationships between entities and creating questions that capture all of this information. We put a more detailed related work review in Appendix 5.

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 knowledge graph and augment its nodes with respective text and molecular structures. This transforms the base knowledge graph 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 knowledge graph, we use an LLM to augment the graph (§3.3), adding another source for question generation. After enhancing our knowledge graph, 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. We explain each step in detail below.

Figure 2: Our dataset development pipeline.

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) [81, 89], which gives insight into the components potentially responsible for participating in various bio-molecular interactions. We provide details of each modality below.

3.1.1 Graph Modality

Our base knowledge graph is from [93], which is used to train a graph neural network for predicting polypharmacy side-effects. 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 drug-drug edges are labeled by their polypharmacy side-effects (~1.3K unique labels) mined from various medical databases.

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. 1) a subgraph with statistics as shown in Table 1.

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 Statistics

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 [58] such as `3351` (`HTR1B`, gene name corresponding to the ID, but mappable to its protein). These are standard 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 [37] and typically not found in research studies on *Fluoxetine*. Moreover, we need to identify the names of those drugs and proteins for searching literature (§3.1.2); while using IDs results in no documents. Specifically, we obtain entity names by querying various databases such as PubChem and STRING [71], depending on which one can provide a generic name. Appendix 4 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) [39], a database integrating information on small molecules (drugs) and their associations with proteins. For each drug-protein pair (c.f. Table 1), we query STITCH to retrieve their interaction, which is provided in two fields, i.e., *mode* (nature of the interaction; *binding*, etc.) and *action* (effect of the interaction; *activation* or *inhibition*). We combine the two fields to label a DPI, which results in two edge types, i.e., *binding and activation* and *binding and inhibition*. Further details in App. 3.

3.1.2 Background Information (Text Modality)

The second modality we deal with is text. To associate each node with a document, we use Wikipedia summaries. We also consider sourcing texts (abstracts) from the PubMed database [82]. However, the texts are usually clinical studies about a *specific* aspect of a drug/protein as opposed to broader knowledge. Additionally, the texts 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 we explain in the Appendix. 4.1, 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 [70]. As such, we limit our search space to those drugs that either have a generic or commercial name identified in Section 3.1.1 to make it easier to obtain background data. This excludes drugs such as `(+)-Vinblastine` (having stereochemistry information) which leaves a total of 198 drugs and 51 proteins with Wikipedia entries. These entities yield 500 DDIs and 84 DPIs as shown in Table 3.

3.1.3 SMILES (Molecular Structure Modality)

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 [63] 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) (Hydrogen atoms are assumed and not shown in the string) and the = refers to a double bond between the carbon atoms. Using these rules, various drugs and compounds are standardized. Proteins do not have a SMILES representation as they are not considered "small molecules" like drugs [44]. 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 [37], a database for studying chemical compounds. Further details on SMILES is provided in App. 2.

3.2 Text Post-Processing (Stage II)

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 (drug discovery date, etc.). Additionally, many summaries are quite verbose, which could potentially inundate an LLM's context window, leading to a *lost-in-the-middle* effect [46], i.e., ignoring information in the middle of the prompt while focusing on the start/end. To address these issues, we utilize GPT's abilities in paraphrasing text [25]. First, we identify documents having more than 200 tokens (a threshold we set empirically). This excludes all protein texts as they have an average of 114 tokens, while for drugs it is 212. Next, we prompt an LLM (GPT-4o [29]) to transform the provided input, i.e., rewriting the source data 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) as it is known that most LLMs have extensive training on Wikipedia [34], we attempt to reduce their chance of *cheating* by using direct references from Wikipedia. Fig. 2 in Appendix gives the prompt we used.

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 adopt `Gunning-Fog` Index [21], a popular score indicating the level of expertise one requires to comprehend a given text. Higher values imply increased complexity suitable for audiences with formal training in a discipline. Results from the process are given in Table 2. As we can see, the enhanced texts are more condensed and their readability scores are almost doubled. This achieves our goal of having a semantically dense corpus that addresses the aforementioned issues, helping create complex queries capable of challenging frontier LLMs. A snippet from post-processing *Fenofibrate*'s (drug) text is given in Figure 3, which shows how simple phrases such as *abdominal pain* get modified to *gastrointestinal disturbances*, etc.

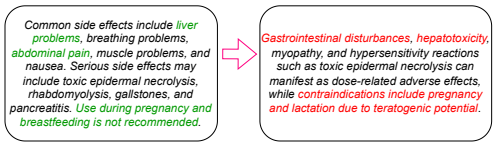

Figure 3: Simple phrases as *liver problems* are replaced by technical jargon as *hepatotoxicity*.

	Before PP	After PP
Tokens	300	233
Readability	15	28

Table 2: Outcome of text before and after post-processing (PP). Each number represents the average over the documents.

3.3 Molecular Interaction Extraction (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 exist between them. Unfortunately, molecular relationships are not provided by the base graph and ignoring them would lead to loosing valuable 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 [23, 26], LLMs have strong capabilities in reasoning through chemistry tasks, including molecular property prediction from SMILES analysis [23]. 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. 3) 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.

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

Table 3: Question distribution in the dataset.

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 revealed five categories of relationships - `Hydrogen Bonding`, π-π `Stacking`, `Steric Clashes`, `Ionic Interaction` and `Electrostatic Interaction`. The introduction of molecular 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 to our dataset. Fig. 4 depicts a portion of the final graph.

3.4 Question Generation (Stage IV)

LLMs have demonstrated great performance in rewriting or generating desired questions that follow a given prompt and context [2, 83, 27, 45]. Leveraging these capabilities, we *prompt* GPT-4.1 [60] for our question generation (prompt in Fig. 5). 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 [9]. 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 knowledge graph triple) and associated background data needed to answer it.

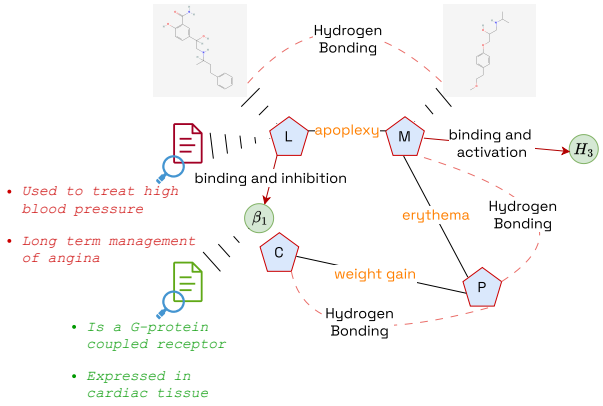

Figure 4: The final knowledge-graph depicting each modality, edge direction and augmented edges (molecular interactions). Entities in blue are drugs, and those in green are proteins. Dashed edges indicate molecular interactions, and solid edges are DDI/DPI, depending on the type of entities. L = *Labetalol*,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

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 [1, 10] 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 for a DPI. To craft questions, the LLM 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 that the LLM (GPT-4.1) must obey when crafting questions: (i) **Entity names must not appear verbatim in the question** - To push a model to understand the

context, connect it to information in the knowledge base and then infer the entity name as opposed to directly learning the entities involved. This in turn 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. Although it is difficult to automatically evaluate the other questions due to variations in phrasing, holistic examination indicates that all questions incorporate a relational component from the knowledge-graph.

We use four quantitative metrics to analyze our questions, i.e., 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). More details of the metrics are given in Appendix 13. From the results in Table 4, we observe that each score reflects the linguistic and semantic complexities proposed by our dataset.

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: Statistics for questions.

3.5 Question Verification (Stage V)

To assess the quality of generated questions, we adopt two methods: (i) **Automatic Verification**. Using LLMs to assess data quality (`LLM-as-a-judge`) [18, 41, 42] has emerged as a promising proxy for human evaluation. By describing a set of grading criteria, or rubric, in the prompt, frontier LLMs such as `GPT` [4, 29] and `Claude` [7] can provide assessments aligning with real annotators [18, 41, 42]. However, using the same LLM for generation and evaluation is inadvisable as they have a tendency to be biased towards their own generations[62]. Hence, we use `Claude-3.7 Sonnet` for automatic evaluation; and (ii) **Human Evaluation**. Despite the proficiency of the `LLM-as-a-judge` setup, they are prone to issues such as preferring verbose outputs over shorter ones and sensitivity to the instructions in the prompt [38]. Thus, having a human-in-the-loop alongside the judge-LLM is an effective strategy to gauge data quality. One of our co-authors, a PhD student in Bioinformatics provides the domain-knowledge required to evaluate the questions.

We randomly sample 100 QA pairs for the LLM and human annotator to provide feedback on. Both are given a rubric (Fig. 4) consisting of 4 custom metrics, i.e., *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 [35]). Each metric is scored as 0, 1 or 2, depending on the guidelines (Fig. 4). Averaged scores for each metric for human and LLM evaluation are given in Table 5. As we can 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 indicate increased complexity in our rubric). This makes sense as `Claude-3.7 Sonnet` is a generalist model [20] whereas our human annotator is a well-read student in the medical domain, who is more comfortable with the language and content of 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.

4 Experiments

We perform two categories of experiments, i.e., *LLM-reasoning* and *retrieval*-based generation, which aims to show the quality and value of `BioMol-MQA` and how existing LLMs perform on it.

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)

4.1 LLM Reasoning Capability and Quality of `BioMol-MQA`

In this section, we investigate LLMs' reasoning capabilities, i.e., how well LLMs perform on our questions without incorporating any sort of `RAG` and how well LLMs perform when they are given the gold data (KG-triples, drug/protein background text and drug SMILES (when applicable)). Through these tests, we determine LLM's performance lower-bound (*zero-shot* knowledge) and upper-bound (directly providing the necessary gold-data), which also shows the quality of our data.

LLMs Used We benchmark seven models on the test split of our dataset, which include: (i) Two closed-source LLMs, citeanthropicClaudeSystem and `o4-mini` [61]; and (ii) five open-source LLMs, `TxGemma` [76] (trained to understand information on various areas of drug development), `DeepSeek-R1` [22], `Llama 3.3` [55], `Qwen3` [73], and `Mistral` [31]. 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 [77, 87]. For open-source LLMs, we run tests with the best available variant attainable via *Together AI* [5].

Metrics LLM answers are evaluated using (i) Lexical EM (exact match): 0 (miss)/1 (hit) measure of equality between two strings, (ii) Lexical F1: Balanced proportion of token overlap between the predicted and reference strings and (iii) `BERTScore` F1 [90]: An embedding-based similarity measure between two strings. By using these three metrics covering different aspects, i.e., exact matching, overlapping of response with groundtruth, and semantic similarity, we can have a better understanding of LLM reasoning capabilities and our dataset quality.

Analysis The results are given in Table 6. 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 Performance of Retriever and Multimodal RAG

The second set of experiments is meant to explore the capabilities of existing retrievers in relation to our dataset. In other words, we want to see how well current retrieval models query and access our knowledge sources. Given the multi-modal nature of our dataset, we investigate three types of retrievers, text-only, graph-only and a combination of both as our multi-modal retrieval baseline which simply combines the results from the best text and graph retriever.

Retrievers Used For text-only retrieval, we test two types of embeddings, i.e., *sparse* - `BM25` [64] and *dense* - [`MedCPT` [33], `DPR` [36], `MolLM` [72], OpenAI's `text-embedding-3-large` (TE3L) [59]]. `BM25` is found to be a competitive baseline [51], occasionally surpassing dense retrievers [51]. `MedCPT`, `DPR` and `MolLM` are BERT-style [13] encoders, which are trained for different purposes such as molecular data understanding [72] and QA [36] tasks. TE3L is a commercial retriever [59]. Details on each model are in Appendix 11.

For our graph-only retriever, we implement a simple baseline using the `Neo4J` database [57]. It works by retrieving knowledge-graph triples semantically similar to the query. The query and triples are encoded using `all-MiniLM-L6-v2` sentence embeddings [66] (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 retrieval. We provide results using the entire constructed graph (Table 1), graph neural network (GNN) based retrievers and, further details on Neo4j in App. 12.

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
`text-embedding-3-large`	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 [BM25 + Neo4j]	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.

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 given in Table 7. We observe: **(i)** For text-only retrievers, we find that a simple BM25 baseline outperforms *each* dense embedding model. This is supported by studies [11, 65] which show how frequency-based 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. 12). As we explain in §App. 12, training GNNs on our graph *directly* is difficult due to its small size (KG's typically have millions of nodes and edges [30]) 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 rate goes down, the 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. 10). 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.

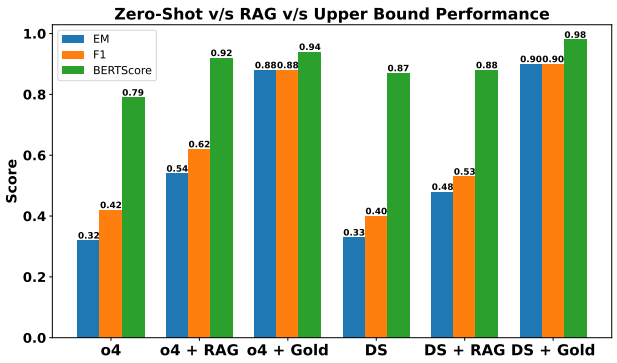

Figure 5: Impact of RAG on o4-mini and DeepSeek R1 (DS). For comparison, their zero-shot and upper bound (with the Gold background data) performances are provided.

5 Conclusion and Future Work

In this paper, we present `BioMol-MQA`, a multi-modal reasoning and retrieval benchmark for polypharmacy queries. Through our experiments, we have shown 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 grow the dataset, such as including *unanswerable* questions, exploring *additional modalities* like protein structures (long-chain amino acid sequence [48]) for testing the length constraints of LLMs [78], and scaling up sample size and considering multi-agent workflows [56] to enhance question diversity.

References

[1] Zahra Abbasiantaeb and Saeedeh Momtazi. 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, 2021.

[2] Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. Synthetic dialogue dataset generation using LLM agents. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.16/.

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

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

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

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

[7] Anthropic. Claude 3.7 system card - Anthropic — docs.anthropic.com, 2025. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf. [Accessed 29-04-2025].

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

[9] Nishant Balepur, Feng Gu, Abhilasha Ravichander, Shi Feng, Jordan Lee Boyd-Graber, and Rachel Rudinger. Reverse question answering: Can an LLM write a question so hard (or bad) that it can't answer? In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. URL https://aclanthology.org/2025.naacl-short.5/.

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

[11] Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.19. URL https://aclanthology.org/2022.findings-emnlp.19/.

[12] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=wS7PxDjy6m.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

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

[15] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 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, 2024.

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

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

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

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

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

[21] Robert Gunning. The fog index after twenty years. *Journal of Business Communication*, 6 (2):3–13, 1969. doi: 10.1177/002194366900600202. URL https://doi.org/10.1177/002194366900600202.

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

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

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

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

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

[27] Hamed Hematian Hemati and Hamid Beigy. Consistency training by synthetic question generation for conversational question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 630–639, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.57. URL https://aclanthology.org/2024.acl-short.57/.

[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

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

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

[31] 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. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[32] Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576, 2024. doi: 10.48550/ARXIV.2405.13576. URL https://doi.org/10.48550/arXiv.2405.13576.

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

[34] Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. Wikimedia data for AI: a review of wikimedia datasets for NLP tasks and AI-assisted editing. In Lucie Lucie-Aimée, Angela Fan, Tajuddeen Gwadabe, Isaac Johnson, Fabio Petroni, and Daniel van Strien, editors, *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 91–101, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wikinlp-1.14. URL https://aclanthology.org/2024.wikinlp-1.14/.

[35] Mael Jullien, Marco Valentino, and André Freitas. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.semeval-1.271. URL https://aclanthology.org/2024.semeval-1.271/.

[36] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

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

[38] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre

Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

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

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

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

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

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

[44] 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, et al. Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9): 1523–1531, 2019.

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

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

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

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

[49] Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 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, 2023.

[50] Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. MoleculeQA: A dataset to evaluate factual accuracy in molecular comprehension. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3769–3789, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.216. URL https://aclanthology.org/2024.findings-emnlp.216/.

[51] Man Luo, Shashank Jain, Anchit Gupta, Arash Einolghozati, Barlas Oguz, Debojeet Chatterjee, Xilun Chen, Chitta Baral, and Peyman Heidari. A study on the efficiency and generalization of light hybrid retrievers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1617–1626, Toronto, Canada, July

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.139. URL https://aclanthology.org/2023.acl-short.139/.

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

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

[54] Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 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, 2025.

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

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

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

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

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

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

[61] OpenAI. OpenAI o3 and o4-mini System Card — openai.com. https://openai.com/index/o3-o4-mini-system-card/, 2025. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. [Accessed 01-05-2025].

[62] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

[63] LibreTexts project. Line Notation (SMILES and InChI), aug 11 2020. URL https://chem.libretexts.org/Courses/Fordham_University/Chem1102%3A_Drug_Discovery_-_From_the_Laboratory_to_the_Clinic/05%3A_Organic_Molecules/5.08%3A_Line_Notation_(SMILES_and_InChI). [Online; accessed 2025-04-23].

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

[65] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.496. URL https://aclanthology.org/2021.emnlp-main.496/.

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

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

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

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

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

[71] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. 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, 2023.

[72] Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. MolLM: a unified language model for integrating biomedical text with 2D and 3D molecular representations. *Bioinformatics*, 40(Supplement_1):i357–i368, 06 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae260. URL https://doi.org/10.1093/bioinformatics/btae260.

[73] Qwen Team. Qwen3, April 2025. URL https://qwenlm.github.io/blog/qwen3/.

[74] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

[75] Caterina Vicidomini and Giovanni N Roviello. Protein-targeting drug discovery, 2023.

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

[77] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

[78] Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 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, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/917. URL https://doi.org/10.24963/ijcai.2024/917.

[79] Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vDFA1tpuLvk.

[80] Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and RISHITA ANUBHAI. Biobridge: Bridging biomedical foundation models via knowledge graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jJCeMiwHdH.

[81] David Weininger. 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, 1988.

[82] Wikipedia contributors. Pubmed — Wikipedia, the free encyclopedia, 2025. URL https://en.wikipedia.org/w/index.php?title=PubMed&oldid=1289629349. [Online; accessed 13-May-2025].

[83] Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. Synthetic multimodal question generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.759. URL https://aclanthology.org/2024.findings-emnlp.759/.

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

[85] Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Empower large language model to perform better on industrial domain-specific question answering. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.29. URL https://aclanthology.org/2023.emnlp-industry.29/.

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

[87] Dharunish Yugeswardeenoo, Kevin Zhu, and Sean O'Brien. Question-analysis prompting improves LLM performance in reasoning tasks. In Xiyan Fu and Eve Fleisig, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 402–413, Bangkok, Thailand, August 2024. Association for Computational Linguistics. ISBN 979-8-89176-097-4. doi: 10.18653/v1/2024.acl-srw.45. URL https://aclanthology.org/2024.acl-srw.45/.

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

[89] 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, Huabin Xing, and Huajun Chen. Scientific large language models: A survey on biological & chemical domains. *ACM Comput. Surv.*, 57(6), February 2025. ISSN 0360-0300. doi: 10.1145/3715318. URL https://doi.org/10.1145/3715318.

[90] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

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

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

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

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our dataset in detail and showcase how existing LLMs, Retrievers and RAG frameworks face challenges when dealing with our dataset (c.f. 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We provide limitations of our work in Appendix 15.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not involve any theoretical work. As such, there are no theorems or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We do not train our LLMs and rely only on out-of-the-box settings as provided by APIs or frozen checkpoints (c.f. Appendix 14). As such, all of our results are reproducible. In the case that we do train models (GNN-retrievers), we provide the training architecture and hyperparameters in Appendix 12.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The first page of the paper provides links to both our code (`https://github.com/saptarshi059/biomolqa`) and dataset (`https://huggingface.co/datasets/BioMolMQA/BioMolMQA`).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our training/validation/test splits in section 3.4. Wherever our paper involves training models (Appendix 12) we describe the hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We run each model once via their API to manage costs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details of our hardware in Appendix 14.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read the code of ethics and found nothing in our work to go against it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction (§1) explains the potential positive effects of curating our dataset, i.e., in evaluating LLM performance for a serious real-world issue (polypharmacy). However, we do not see any malicious effect that can stem from our dataset.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our dataset does not pose such risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The authors of the relevant methods, including the source of our knowledge graph have been properly cited throughout our work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The first page of the paper provides links to both our code (`https://github.com/saptarshi059/biomolqa`) and dataset (`https://huggingface.co/datasets/BioMolMQA/BioMolMQA`).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our research does not involve human subjects. We only have an evaluator who was provided instructions for evaluating our dataset (Fig. 4, c.f. 3.5).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research does not involve human subjects. We only have an evaluator who was provided instructions for evaluating our dataset (Fig. 4, c.f. 3.5).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The use of LLMs in our work for synthetic data generation is completely explained throughout the dataset construction section (c.f. 3).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.