From Graphs to Questions: A Framework for Complex Biomedical KGQA Dataset Generation

Anonymous ACL submission

Abstract

This work introduces BioGraphletQA, a novel large-scale dataset for complex biomedical Knowledge Graph Question Answering (KGQA) and describes the underlying generation framework. Central to our approach is the use of graphlets—small subgraphs extracted from a KG—as anchors for generating diverse and complex QA pairs using large language models (LLMs). Our pipeline comprises three stages: (1) KG preprocessing and reduction to produce a manageable subset; (2) an extensive prompt ablation study to identify the optimal prompt for QA generation; and (3) a filtering phase using an LLM to refine the dataset by removing low-quality pairs. The final dataset comprises 119,856 complex QA pairs, each linked to a graphlet containing up to five nodes. To assess quality, a domain expert annotated 53 QA pairs across five criteria, confirming the scientific validity, complexity, and completeness of the data. All code is available at: https://anonymous.4open.science/r/Synthetic-KGQA-CE2F.

1 Introduction

Question answering (QA) systems have benefited immensely from advances in large language models (LLMs), particularly Transformer-based architectures (Vaswani et al., 2017). However, despite their success, LLMs struggle with factual consistency, often generating hallucinated or inaccurate responses (Ji et al., 2023; Huang et al., 2025). One promising approach to mitigate these issues is the use of Knowledge Graph Question Answering (KGQA) datasets. Traditional KGQA datasets, however, are either manually curated—making them costly and time-intensive (Gu et al., 2021)—or template-based, which often limits their diversity and generalizability (Banerjee et al., 2023).

In the biomedical domain, the problem of hallucinations can lead to dangerous outcomes such as misdiagnoses, unsafe treatment recommendations, and compromised patient safety. Although several biomedical KGs exist—such as OREGANO KG (Boudin et al., 2023), CKG (Santos et al., 2022), MonarchKG (Putman et al., 2023), and PrimeKG (Chandak et al., 2023)—most KGQA research has focused on large open-domain KGs like Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014), which often lack the granularity and reliability required for biomedical decision-making. To date, only one large-scale synthetic biomedical KGQA dataset has been developed (Yan et al., 2024), generated using an LLM with graphlets from PrimeKG, underscoring the need for more robust, domain-specific QA resources.

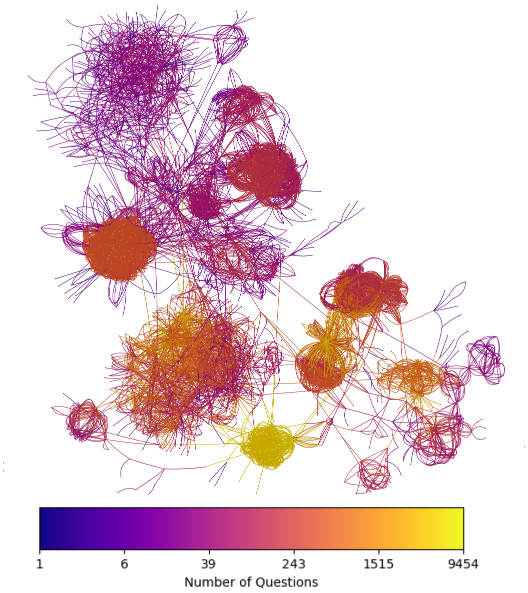


Figure 1: Weighted sample of the KG of our dataset. 2,000 (out of 18,954) nodes were sampled. Node color indicates the number of questions generated per node.

In this work, we propose a novel multi-stage method to generate a large-scale, high-quality, complex biomedical KGQA dataset applied to the biomedical domain. Our approach includes an ablation study to identify the best prompt for generating the initial dataset, followed by a post-generation filtering stage using an LLM to enhance dataset quality. For generation, we feed the graphlet as plain text and give the LLM freedom to select which nodes in the graphlet are most relevant to generate the QA pair. This is opposed to template based methods which force the use of all the nodes in a pre-defined manner. The resulting dataset contains 119,856 filtered QA pairs, each linked to a graphlet from the OREGANO KG (v2.1) (Boudin et al., 2023), with QA pairs spanning 29 different graphlet shapes ranging from 3-5 nodes, offering a wide variety of questions.

Our methodology is divided into three main stages: KG cleaning, where we hydrate and reduce the size of the KG as well as generate the graphlets; an initial dataset generation stage, informed by the prompt ablation study; and automatic filtering, which removes around half of the generated dataset. A small sample of the number of questions generated per node is shown in Figure 1. Finally, we perform human evaluation, by a domain expert, on a small sample to verify the quality of the generated data. This work was inspired by the following research questions.

RQ1: How to utilize KGs to effectively generate question-answer pairs?

RQ2: How can we systematically compare and evaluate prompts to optimize QA generation?

RQ3: How can we assess the quality and reliability of synthetically generated QA pairs?

With this work, we make two main contributions to the fields of QA and bioinformatics. First, we introduce BioGraphletQA, a large-scale biomedical KGQA dataset designed to support the training and evaluation of future KGQA systems. Second, while we demonstrate a use case in the biomedical domain, we present a data-agnostic generation pipeline that can be applied to other KGs, enabling the scalable construction of complex synthetic KGQA datasets across diverse domains. Beyond its role in KGQA research, BioGraphletQA also serves as a standalone resource for complex biomedical QA.

2 Related Work

Recent advances in LLMs have spurred a surge in using synthetic data generation to overcome data scarcity and privacy challenges in IR and QA tasks. For instance, Braga et al. (2024) propose a framework that generates synthetic answers tailored for personalized community QA, demonstrating that fine-tuning on this generated data can yield performance comparable to models trained on human-curated datasets. Similarly, Tang et al. (2023) explores leveraging ChatGPT to generate synthetic clinical documents, reporting substantial improvements in downstream tasks like named entity recognition and relation extraction. In addition, GeMQuAD, introduced by Namboori et al. (2024), employs few-shot learning with LLMs to create multilingual QA datasets, thereby enhancing performance in low-resource settings. Complementing these efforts, Wu et al. (2024) present a synthetic multimodal question generation approach that combines the strengths of LLMs and multimodal models to produce high-quality QA pairs from diverse document types.

KGQA datasets have evolved significantly, with several notable benchmarks such as LC-QuAD (Dubey et al., 2019) and ComplexQuestions (Bao et al., 2016). GrailQA (Gu et al., 2021) and GrailQA++ (Dutt et al., 2023) advanced the field by introducing a dataset specifically designed to evaluate generalization in KGQA systems across different levels of compositional complexity. Jiang and Usbeck (2022) provided a comprehensive survey of KGQA methods and datasets, highlighting the challenges and opportunities in this domain.

Recent advances in biomedical KGQA include PrimeKGQA (Yan et al., 2024), which contains approximately 84,000 QA pairs generated through few-shot prompting using graphlets extracted from PrimeKG. This approach builds on graphlet-based methodologies similar to those in GrailQA++ (Dutt et al., 2023). Our work follows a similar graphlet-based idea but extends it by introducing a more in-depth prompt selection strategy and an additional QA filtering phase to improve quality. Similarly, ConvKGYarn (Pradeep et al., 2024) generates synthetic QA pairs by combining KG facts with slot-filled question templates. While this enables large-scale QA generation the reliance on predefined templates can limit question diversity and contextual depth. In contrast, our dynamic node selection strategy allows the LLM to flexibly identify relevant nodes and relations within each graphlet, leading to more varied and contextually nuanced QA generation.

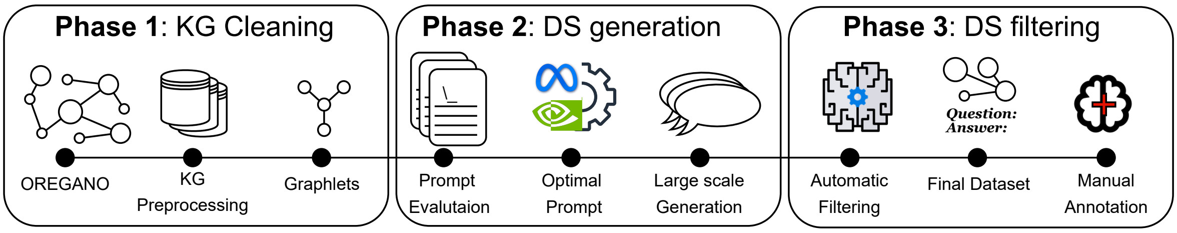


Figure 2: An overview of our methodology, composed of three phases: 1) Initial cleaning of the OREGANO KG, including KG pre-processing and graphlet extraction; 2) Generation of the initial KGQA dataset, starting with a prompt evaluation stage; 3) Automatic filtering stage using an LLM-as-a-judge, with human evaluation.

3 Methodology

In this section we present the methodology used in this work with the aim of creating and evaluating a KGQA dataset where each QA pair is linked to a graphlet representing facts from the KG. Our methodology uses a variety of graphlet shapes to build questions of different complexity. An overview of our method can be seen in Figure 2. First (3.1), we discuss and justify the choice of the KG used in this work and describe the process of hydrating the names for the entities in the graph, the graph reduction techniques used, and finally the process for graphlet extraction. Following this, we discuss the process used to generate our dataset (3.2), especially focusing on an ablation study to determine the best structure for the prompt. We then discuss the post generation filtering techniques used to improve the quality of the dataset, as well as the process of manual evaluation (3.3).

3.1 Knowledge Graph

To develop a robust KGQA dataset, selecting an appropriate KG is crucial. While our approach is adaptable across domains, the biomedical field offers unique challenges and opportunities. From the numerous existing biomedical KGs (Haas, 2024), we sought one that balances size and complexity, ensuring diverse node classes linked to reputable biomedical databases for comprehensive question generation. Based on these criteria, we selected the OREGANO KG (v2.1) (Boudin et al., 2023), which contains 88,937 nodes spanning 11 types[1] and 824,231 edges with 19 edge types.

3.1.1 Hydration

Whilst the OREGANO dataset includes entity names for some of its nodes, we still needed to hydrate certain nodes, and we opted to update most data where possible. As a result, we had to look up various identifying terms. Each identifying term was looked up between December 3 and 19, 2024. Furthermore, we ensured that the licenses for all the knowledge bases allowed us to publish the names accordingly. The preferred order of identifiers for each entity class is provided in Appendix B. This led to a total of 85,655 denormalized nodes, of which 81,240 (94.85%) are unique.[2]

3.1.2 Reduction

After hydrating the KG, we analyzed its node degree distribution. As shown in Figure 3, a substantial number of nodes have a degree of one (edge nodes) while a small subset exhibit very high degrees (hub nodes). We hypothesize that the edge nodes offer limited value since they are in general associated with few nodes, reducing variability of questions, and the hub nodes risk redundancy by appearing in too many questions. According to this hypothesis, we filter nodes with a degree greater than 100 or less than 3. This reduction is intended to enhance the variability of nodes in our dataset while preserving meaningful structural complexity, as well as reducing the size of the KG, making further processing easier. Following this reduction, the graph comprises 41,115 nodes and 129,992 edges, with the updated node degree distribution shown in Figure 3, and node type distribution illustrated in Appendix C (Figure 10).

3.1.3 Graphlets

The final pre-processing step involves extracting graphlets. Instead of performing a simple random walk, we partition the KG into graphlets and use these substructures as the foundation for generation. Graphlets are small, connected, non-isomorphic subgraphs that encapsulate local structural patterns within a larger graph or network. In our

[1] Note that there is also a ‘code’ entity class, which we did not utilize.

[2] Two entities were not hydratable: one disease and one pathway

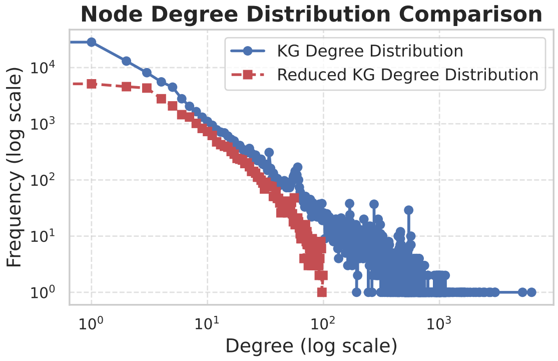


Figure 3: The node degree distribution of the OREGANO KG and the reduced version.

approach, we consider all 29 unique graphlet structures containing between 3 and 5 nodes (2 being trivial). These structures are illustrated in Figure 5. Subgraph enumeration is generally computationally expensive (Ribeiro et al., 2021), making our graph reduction techniques particularly advantageous. To efficiently identify graphlets, we employ the `graph-tool` library (Peixoto, 2014).

Our method involves several key steps. First, the input data is loaded into `graph-tool` as an undirected graph that is then preprocessed to remove parallel edges. We count the frequency of each graphlet shape using `gt.motifs()`, and apply sampling following the approach of Wernicke (2006) to target approximately 10,000 graphlet occurrences for each shape, to control dataset size (the most frequent graphlet appears over 1.8 trillion times).

This process results in a final dataset of 269,574 graphlets, which serves as the foundation for generating our initial QA dataset. All relevant statistics can be seen in Table 1 in the Appendix.

3.2 Dataset Generation

With the graphlet extraction complete, we proceed to large-scale KGQA dataset generation. A key preliminary step is selecting an effective prompt. In our approach, each graphlet inherently encodes both the question and answer. The question targets one or two *Question Nodes*, while *Hidden Nodes* facilitate reasoning to infer the *Answer Node*. To provide structured graph representation, we explicitly include the graphlet's shape (edge list) and node names in the prompt. We omit the edge type, as letting the LLM infer the relation yields better results than providing a simple type like `has_effect`.

3.2.1 Prompt Ablation

When working with LLMs, minor prompt modifications can significantly affect performance (White et al., 2023; Sclar et al., 2023). While newer, more powerful models have somewhat mitigated this issue, prompt engineering remains an important factor. To address this, we conducted an ablation study to evaluate various prompts and identify the most effective one for our dataset generation.

Although automated approaches, such as gradient-based optimization (Pryzant et al., 2023; Chen et al., 2023) and gradient-free alternatives (Prasad et al., 2022) exist for this task, they typically rely on ground truth evaluations. Since our dataset is fully synthetic, we could not provide definitive ground truth examples. Instead, inspired by the use of LLMs-as-evaluators (Li et al., 2024; Bai et al., 2023; Gao et al., 2024) and LLM-as-a-judge (Zheng et al., 2023), we used an LLM (Llama-3.1-Nemotron-70B) to score the generated instances considering a set of characteristics identified during initial prompt testing. Each characteristic is encoded as a prompt, presented in Appendix D.1, and contributes with 1 point for a possible maximum score of 6:

1. **Answer node present in question:** Ensures the question does not contain the answer in an obvious way. Example: *"Is X a side effect of drug Y?"* (This would be a simple yes/no question).

2. **Question contains graphlet-based terminology or hints:** Prevents the generated questions from explicitly referencing the graph structure. A common issue was that questions mentioned "connections" between entities, which is not typical of language used by biomedical experts. Example: *"What is the connection between X and Y?"*

3. **Answer contains graphlet-based terminology:** Similar to the previous feature, but focused on ensuring that answers do not explicitly describe connections between entities.

4. **Scientifically accurate question:** Ensures that the generated question is meaningful and logically sound from a biomedical perspective.

5. **Scientifically accurate answer:** Ensures that the provided answer is scientifically valid and free from inaccuracies.

6. **Question is properly answered:** Verifies that the answer correctly addresses the question without ambiguity or irrelevance.

For the ablation study, we adopted a modular approach to design 15 distinct prompt variations, each combining different subsets of prompt components. These prompts were evaluated on a dataset of 1,000 randomly selected graphlets. Inspired by Chain of Thought (CoT) prompting (Wei et al., 2022), we provided the model with structured instructions designed to guide its output toward components useful for the final answer. This is similar to CoT prompting however rather than the model thinking step-by-step, we give the reasoning steps it should follow. We also tested a reflection module, which asked the model to critique and refine its initial responses. The 15 prompt configurations can be grouped into five categories, as described below:

1. **Baseline Prompts**: Targeted at setting baselines.

    1.1 `[Baseline]`: The simplest version of the prompt with no additional instructions or examples.

    1.2 `[1.1 + Simple Example]`: The baseline prompt with a simple example to guide the model.

2. **QA Instruction Prompts**: Gives the model strict instructions on how to generate QA pairs. Every time an instruction is given, the prompt uses an additional `Instruction Markdown` to format the instructions properly.

    2.1 `[1.1 + Question Instruction + Answer Instruction]`: The baseline prompt with additional structured instructions on generating questions and answers.

    2.2 `[2.1 + Simple Example]`

3. **Graphlet Analysis Prompts**: Here we try to force the model to analyze the graphlet.

    3.1 `[1.1 + Analyze Graphlet Instruction + Final Analysis]`

    3.2 `[3.1 + Node Types]`: Asks the model to find *Question, Answer* and *Hidden Nodes*.

    3.3 `[3.2 + Simple Example]`

    3.4 `[3.2 + QA Instructions]`

    3.5 `[3.4 + Simple Example]`

4. **Reflection Prompts**: Get the model to reflect on its generated QA pair and improve it.

    4.1 `[1.1 + Reflection Instruction]`

    4.2 `[4.1 + Complex Example]`: Adds a complex example that includes graphlet analysis, reflection and rewriting of the QA.

    4.3 `[4.1 + Question and Answer Evaluation]`: Adds explicit criteria.

    4.4 `[4.3 + Complex Example]`

    4.5 `[4.4 + QA Instruction]`

5. **Full Prompt (All Modules)**: A comprehensive prompt that integrates all components into a single structured format. See Figure 4.

Before extracting features, we first ensured that the generated QA pairs were **JSON-parsable**, a critical requirement. If a prompt produced an output that failed JSON parsing, it was automatically assigned a score of zero for that specific graphlet, as prompts which do not generate valid JSON, should be negatively penalized. The final prompt scores were then calculated as the average across all tested graphlets. Our baseline prompt (Prompt 1.1) achieved a score of `3.45/6`, while our best-performing prompt, Prompt 5 (Full Prompt), scored `4.91/6`. Figure 4 presents Prompt 5 along with most of the modules used.

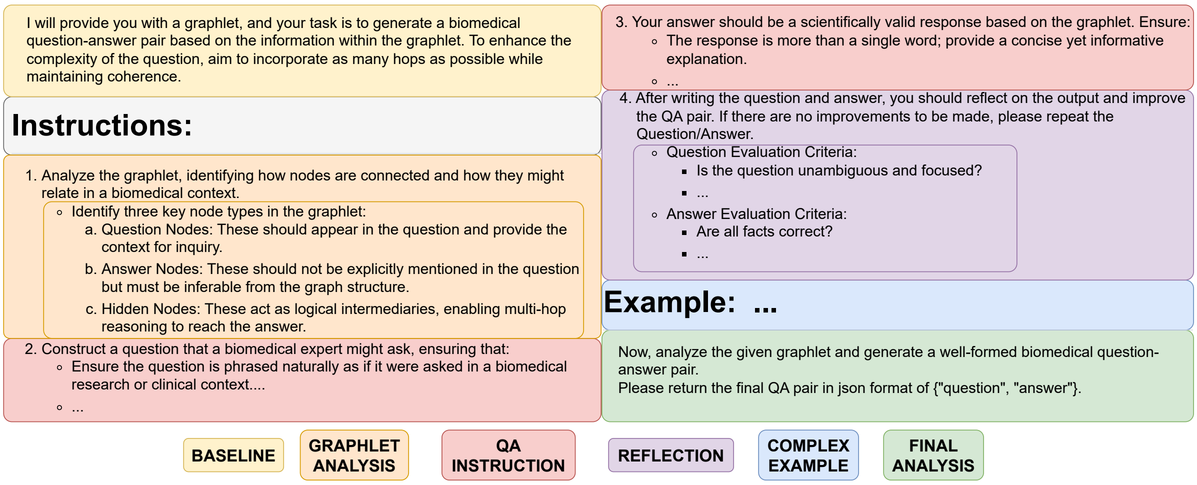
3.2.2 Large Scale Generation

With both the graphlets and the optimal prompt selected, large-scale dataset generation was subsequently performed. This was conducted on the server specified in Appendix A, utilizing the LMDeploy package. At the time of writing, LMDeploy (LMDeploy Contributors, 2023) was among the fastest library for LLM inference. Specifically, we used an AWQ 4-bit quantized version of Llama-Nemotron-70B[3], converted for TurboMind, for all dataset generation (specific model configuration detailed in Appendix D). At the time of testing, this was one of the best open-source models we could run locally[4] (Adler et al., 2024; Wang et al., 2024).

The generation process took just under 10 days, with an average throughput of 27,821 questions per day (241 tokens per second), resulting in a total of 269,574 questions. We do not use 543 outputs since they were not JSON-parsable. We then performed

[3] Nvidia-Llama-3.1 model on Hugging Face

[4] Model selected with LLM Arena (Chiang et al., 2024)



I will provide you with a graphlet, and your task is to generate a biomedical question-answer pair based on the information within the graphlet. To enhance the complexity of the question, aim to incorporate as many hops as possible while maintaining coherence.

Instructions:

1. Analyze the graphlet, identifying how nodes are connected and how they might relate in a biomedical context.
   - Identify three key node types in the graphlet:
     a. Question Nodes: These should appear in the question and provide the context for inquiry.
     b. Answer Nodes: These should not be explicitly mentioned in the question but must be inferable from the graph structure.
     c. Hidden Nodes: These act as logical intermediaries, enabling multi-hop reasoning to reach the answer.
2. Construct a question that a biomedical expert might ask, ensuring that:
   - Ensure the question is phrased naturally as if it were asked in a biomedical research or clinical context....
   - ...
3. Your answer should be a scientifically valid response based on the graphlet. Ensure:
   - The response is more than a single word; provide a concise yet informative explanation.
   - ...
4. After writing the question and answer, you should reflect on the output and improve the QA pair. If there are no improvements to be made, please repeat the Question/Answer.
   - Question Evaluation Criteria:
     - Is the question unambiguous and focused?
     - ...
   - Answer Evaluation Criteria:
     - Are all facts correct?
     - ...

Example: ...

Now, analyze the given graphlet and generate a well-formed biomedical question-answer pair.
Please return the final QA pair in json format of {"question", "answer"}.

Figure 4: Compressed version of Prompt 5, showcasing extracts from some of the different modules. The full prompt and the modules can be seen in our GitHub Repository.

a Z-score analysis, eliminating all QA pairs where either the length of the question or of the answer were outliers (more than three standard deviations from the mean). The acceptable range of the length was `[79, 365]` characters for questions and `[59, 997]` characters for answers. A total of 4,658 QA pairs failed this Z-score test and were removed.

3.3 Post-Generation Filtering

To improve dataset quality, we applied LLM-based filtering after the initial generation phase. Although the prompt used for generation included a reflection phase, several issues could still arise. For example, some graphlets may not contain a valid QA pair worth generating, or the generated answer may be incomplete, requiring additional knowledge to be fully correct. To mitigate these and other potential issues, we applied automatic filtering to the dataset.

The filtering prompt, detailed in Appendix D.3, first instructs the model to evaluate the connections between the entities in the question to determine whether the question is coherent. Next, it attempts to answer the question and compares its response with the previously generated answer. This evaluation is structured in a JSON format to ensure that two boolean variables, `valid_question` and `original_answer_valid`, are generated based on the model's reasoning. While the goal is to use KGs for grounding and reducing hallucination, this step assesses whether the QA pair remains valid based on the LLM's general biomedical knowledge, independent of the specific graph context.

The throughput of the filtering is significantly higher at 45,674 questions per day, taking 6 days to evaluate all the pairs. After filtering, 119,856 QA pairs remained (45% of the dataset after post-processing). Additionally, 17,076 outputs (6.45%) were unparseable as JSON, a notably higher failure rate than during the generation phase. The distribution of the graphlets accepted can be seen in Figure 5. Exact statistics regarding this information is present in Table 1 in Appendix F.

3.3.1 Human Evaluation

Following our automatic filtering process, we conducted human annotation to assess the quality of the dataset. The annotation was performed by the second author, who holds relevant domain expertise in biomedical sciences. We selected two positive QA pairs for each template, totaling 58 QA pairs. We utilized a 5-point Likert-based evaluation criteria (Likert, 1932), with the following categories: Scientific Validity of the Question, Scientific Validity of the Answer, Answer Relevance to the Question, Question Complexity, Specificity of the Answer, Answer Completeness. We also allowed the expert to not rate QA pairs if they lacked confidence. The exact scale is present in Appendix E.

4 Results

In this section, we present an in-depth analysis of our experimental findings, focusing on the impact of different prompt configurations on model performance, as well as presenting some findings about the dataset, and finally presenting the results of the human evaluation.

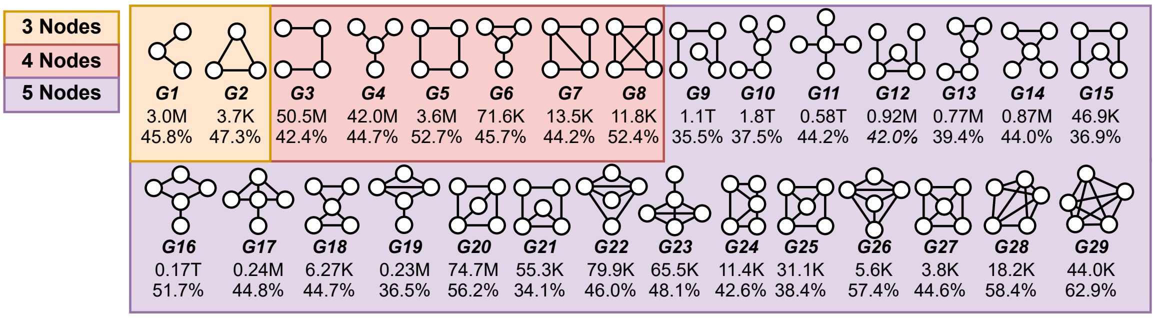


Figure 5: Figure showing the 29 graphlet shapes with 3-5 nodes. Each graphlet has the number of graphlets initially present (sampled to 10,000) and the acceptance ratio (QA pairs accepted / QA pairs generated).

4.1 Prompt Ablation

In analyzing the performance of different prompt configurations, several key insights emerge, as depicted in Figure 6. The baseline prompt, without any additional instructions or examples, yielded a score of 3.45 (1.1). Adding a simple example slightly improved the score to 3.94 (1.2). The inclusion of question and answer instructions did not have any impact; however, with an example, we note an increase to 4.39 (2.2). When exploring graphlet analysis, the results varied. The baseline achieved a score of 4.43 (3.1), but with node type identification, the score dropped to 3.96 (3.2). Adding a simple example to this configuration improved performance to 4.37 (3.3). A similar effect is seen with the addition of QA instructions (4.25, 3.4) and its example (4.40, 3.5).

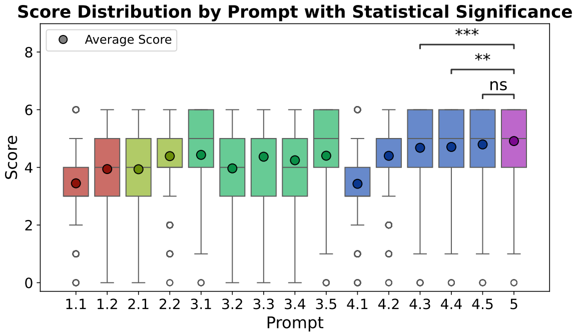


Figure 6: Boxplot of prompt ablation results. The top prompt is statistically compared to the second and third best.

Reflection-based prompt had varying performance with scores of 3.43 for the baseline reflection prompt (4.1). However, adding example-based prompts and evaluation instructions led to a noticeable improvement, with a final score of 4.79 (4.5), demonstrating that reflection combined with evaluation can be beneficial when used alongside examples. Finally, the most complex configuration, achieved the highest score of 4.91 (5). This suggests that while simple configurations provide some improvements, combining various types of instructions and examples yields the best overall performance. According to the Mann–Whitney–Wilcoxon test (Mann and Whitney, 1947), the difference in performance between Prompt 5 and 4.5 is non-significant, however we believe the slightly more complex prompt would perform better.

4.2 The dataset

As discussed, prompt filtering reduced the dataset by approximately 55%, resulting in 119,856 KGQA pairs. After filtering, we reconstructed the KG by selecting the nodes and edges used to generate the filtered dataset. This reconstruction yielded a KG containing 18,954 nodes and 65,015 edges. A sample of this KG, weighted by the number of questions generated per node, is shown in Figure 1, with, a more detailed distribution of the number of questions per node present in Appendix F, Figure 11. From these two figures, we can conclude that a well-distributed number of questions were generated per node, with only 41 nodes generating more than 2,000 questions and a maximum of 9,454 questions generated from a node.

To further explore our earlier hypothesis—that nodes with higher degree tend to generate more questions—we present Figure 7, which compares the ranking of nodes by the number of questions with the ranking by degree. While a correlation can be observed, it is not particularly strong, likely due to the sampling method. This suggests that nodes with high degrees generally correspond to a high number of questions, but the inverse is less consistently true. This phenomenon can be attributed to graphlet structures; for instance, a weakly connected node linked to a highly connected node may appear in many graphlets involving the latter.

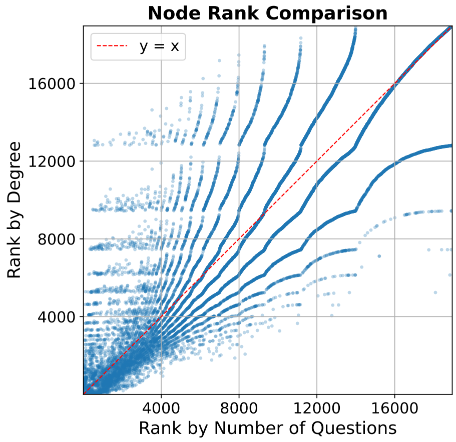


Figure 7: A scatter plot showing the nodes ranked by the number of questions vs nodes ranked by the degree in the reconstructed KG.

4.3 Human Evaluation

Overall, the results of the human annotation process are encouraging and affirm the quality of the dataset. We evaluated a total of 53 QA pairs across five distinct criteria, following the exclusion of five samples due to annotator expertise gaps. The distribution of annotation scores across these criteria is shown in Figure 8. Beginning with the questions, all were rated as scientifically valid. In terms of complexity, the majority (88.68%) received scores of 3 or higher. This aligns well with our aim to generate complex biomedical questions.

Turning to the answers, the overall assessment is similarly positive. On the dimension of scientific validity, 90.57% of responses scored at least a 3, indicating only some minor scientific inaccuracies. Completeness was also a strong point: 92.45% of answers scoring at least 3, suggesting they generally addressed the question with some minor information lacking. However, specificity emerged as a relative weakness. Ideally, answers should receive a score of 3—indicating an appropriate level of detail —but only 73.58% of responses met this benchmark. This suggests that while most answers were correct, some may lack the precision necessary for high-quality biomedical communication.

Further insight comes from analyzing the minimum score across the three answer-related criteria (scientific validity, completeness, specificity). Here, 71.7% of QA pairs achieved a minimum score of at least 3, indicating that the majority of answers were acceptable. More granular statistics, along with representative examples of annotated QA pairs, can be found in Appendix G.

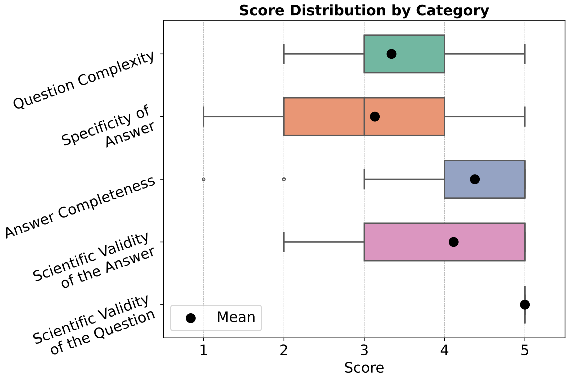


Figure 8: Boxplot of Likert based human evaluation scores across 5 categories.

5 Conclusion

In this work, we introduced BioGraphletQA, a large-scale biomedical KGQA dataset generated from the OREGANO KG through a structured three phase pipeline: KG preprocessing, initial generation, and post-generation filtering. The resulting dataset comprises 119,856 high-quality QA pairs across 29 distinct graphlet types, with validation from a biomedical expert confirming the accuracy and relevance of both questions and answers. This work was driven by several key research questions. Our end-to-end pipeline demonstrates how complex, diverse QA pairs can be systematically generated from a KG (RQ1). To simplify the KG and enhance its utility for QA generation, we employed node reduction techniques that also improved the distribution of generated data. We conducted a thorough prompt ablation study using LLM-as-a-Judge to compare our module based prompts (RQ2). Finally, we evaluated the synthetic QA data through expert review (RQ3), establishing the dataset's high quality and reliability. Beyond its immediate contributions to KGQA, BioGraphletQA offers a novel resource with higher question complexity than existing biomedical QA datasets, supporting more advanced QA models. Moreover, the methods developed here are broadly applicable, providing a scalable and adaptable framework for QA dataset construction across domains. Overall, this work makes contributions to both bioinformatics and the broader QA research community.

6 Limitations

While BioGraphletQA presents a meaningful step forward in large-scale biomedical KGQA dataset generation, several limitations remain:

Lack of Automatic Metrics. In this work, we prioritized human evaluation, over any kind of automatic evaluation mainly due to the lack of reliable metrics to evaluate.

LLM-Induced Biases and Hallucinations. Despite employing an LLM-based filtering step, the generation process is still inherently dependent on the initial LLM's outputs. Hallucinations, biases, or inaccuracies may persist in cases where the filtering model fails to catch them. Moreover, while using a different model for filtering could have been beneficial, no superior open-source LLM was available at the time of writing, and we decided it is more beneficial to filter with a stonger model rather than a different one.

Model Quality Affects Dataset Quality. The overall quality of the generated QA pairs is bound by the capabilities of the LLM used. As better LLMs become available, performance in both generation and filtering could be significantly improved, leading to higher-quality datasets. Further this work uses explicitly open source models, however with closed source models, improvements in the data quality should be observed.

Scalability. Scaling the approach to very large KGs may be hindered by graphlet enumeration and sampling limitations, which is not evaluated in this work, however with enough time and resources this should not be a major limitation.

No Guarantee on Graphlet Utilization. The QA generation process does not enforce the use of all nodes in each graphlet. While this allows for more natural and flexible question construction compared to rigid templates, it also introduces ambiguity about the completeness of graphlet utilization. Some questions may underutilize the full graphlet context, potentially missing the opportunity for deeper graph-based reasoning. This is seen in the 'simpler' questions generated.

Ethical Considerations and Risks

BioGraphletQA is a synthetic dataset generated using LLMs and structured knowledge from the OREGANO biomedical knowledge graph. It is intended exclusively for research purposes in developing and evaluating KGQA systems. The content within the dataset does not constitute medical advice and should not be used to inform clinical decisions or health-related practices.

Acknowledgments

References

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167.

Debayan Banerjee, Sushil Awale, Ricardo Usbeck, and Chris Biemann. 2023. Dblp-quad: A question answering dataset over the dblp scholarly knowledge graph. *arXiv preprint arXiv:2303.13351*.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514, Osaka, Japan. The COLING 2016 Organizing Committee.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Marina Boudin, Gayo Diallo, Martin Drancé, and Fleur Mougin. 2023. The oregano knowledge graph for computational drug repurposing. *Scientific data*, 10(1):871.

Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. Synthetic data generation with large language models for personalized community question answering. *arXiv preprint arXiv:2410.22182*.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*.

Lichang Chen, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. 2023. Instructzero: Efficient instruction optimization for black-box large language models. *arXiv preprint arXiv:2306.03082*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*, pages 69–78. Springer.

Ritam Dutt, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah. 2023. Grailqa++: A challenging zero-shot benchmark for knowledge base question answering. In *Proceedings of the 13th IJCNLP*, pages 897–909.

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. *BMC medical informatics and decision making*, 18:1–19.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Robert Haas. 2024. A survey of biomedical knowledge graphs and of resources for their construction.

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).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Longquan Jiang and Ricardo Usbeck. 2022. Knowledge graph question answering datasets and their generalizability: Are they enough for future research? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3209–3218.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2025. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. 2024. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for nlg evaluation: Advances and challenges. *arXiv preprint arXiv:2401.07103*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

LMDeploy Contributors. 2023. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. 2023. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678.

Amani Namboori, Shivam Mangale, Andy Rosenbaum, and Saleh Soltan. 2024. Gemquad: Generating multilingual question answering datasets from large language models using few shot learning. *arXiv preprint arXiv:2404.09163*.

McKusick-Nathans Institute of Genetic Medicine. 2025. Online mendelian inheritance in man, omim®. Accessed on 25 February 2025.

National Library of Medicine (US). 2024. Umls knowledge sources. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html. Release 2024AA. Bethesda (MD): National Library of Medicine (US); 2024 May 6 [cited 2024 Jul 15]. Available from: http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.

Tiago P. Peixoto. 2014. The graph-tool python library. *figshare*.

Ronak Pradeep, Daniel Lee, Ali Mousavi, Jeffrey Pound, Yisi Sang, Jimmy Lin, Ihab Ilyas, Saloni Potdar, Mostafa Arefiyan, and Yunyao Li. 2024. ConvKGYarn: Spinning configurable and scalable conversational knowledge graph QA datasets with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1176–1206, Miami, Florida, US. Association for Computational Linguistics.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Tim E Putman, Kevin Schaper, Nicolas Matentzoglu, Vincent P Rubinetti, Faisal S Alquaddoomi, Corey Cox, J Harry Caufield, Glass Elsarboukh, Sarah Gehrke, Harshad Hegde, Justin T Reese, Ian Braun, Richard M Bruskiewich, Luca Cappelletti, Seth Carbon, Anita R Caron, Lauren E Chan, Christopher G Chute, Katherina G Cortes, Vinícius De Souza, Tommaso Fontana, Nomi L Harris, Emily L Hartley, Eric Hurwitz, Julius O B Jacobsen, Madan Krishnamurthy, Bryan J Laraway, James A McLaughlin, Julie A McMurry, Sierra A T Moxon, Kathleen R Mullen, Shawn T O'Neil, Kent A Shefchek, Ray Stefancsik, Sabrina Toro, Nicole A Vasilevsky, Ramona L Walls, Patricia L Whetzel, David Osumi-Sutherland, Damian Smedley, Peter N Robinson, Christopher J Mungall, Melissa A Haendel, and Monica C Munoz-Torres. 2023. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52(D1):D938–D949.

Pedro Ribeiro, Pedro Paredes, Miguel EP Silva, David Aparicio, and Fernando Silva. 2021. A survey on subgraph counting: concepts, algorithms, and applications to network motifs and graphlets. *ACM computing surveys (csur)*, 54(2):1–36.

Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. 2022. A knowledge graph to interpret clinical proteomics data. *Nature biotechnology*, 40(5):692–702.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, et al. 2021. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10–D17.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

P Talapova, MA Gargano, N Matentzoglu, B Coleman, EB Addo-Lartey, AV Anagnostopoulos, J Anderton, and P Avillach. 2023. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

The UniProt Consortium. 2024. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

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

Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. 2008. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519.

Sebastian Wernicke. 2006. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4):347–359.

Michelle Whirl-Carrillo, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F Thorn, Ryan Whaley, and Teri E Klein. 2021. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110(3):563–572.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Pakazad, Tongshuang Wu, and Graham Neubig. 2024. Synthetic multimodal question generation. *arXiv preprint arXiv:2407.02233*.

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.

Xian Zeng, Peng Zhang, Weidong He, Chu Qin, Shangying Chen, Lin Tao, Yali Wang, and et al. 2018. Npass: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research*, 46(D1):D1217–D1222.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Implementation Details

All work was conducted on a server that contained an A6000 with 48GB of VRAM, 256GB of RAM and an AMD EPYC 7543 (32C/64T). Due to fair usage of the server, we limited the number of CPU cores to 4 and the RAM of the machine to 16GB. Further each job is limited to 2 day of runtime. Further the execution of graphlet counting, was conducted on a separate server, with 24 cores and 128GB of RAM (<11 hours).

B Knowledge Graph Hydration

One of the disadvantages of working with the OREGANO dataset is the absence of textual names for majority of the nodes in the graph. For example, a node such as COMPOUND:786 is only represented by its corresponding biomedical database identifiers. As a result, we had to look up these various identifiers. Each identifier was looked up between December 3 and 19, 2024. Furthermore, we ensured that the licenses for all the knowledge bases allowed us to publish the names accordingly. The preferred order of identifiers for each entity class is as follows:

- **Compound** (32,083): Already hydrated (5,165), PubChem Compound (24,642) (Kim et al., 2025), DrugBank (Knox et al., 2024) (910), NPASS (1,225) (Zeng et al., 2018), SIDER (103) (Kuhn et al., 2016), PharmGKB (38) (Whirl-Carrillo et al., 2021)
- **Protein** (14,505): UniProtKB (13,355) (The UniProt Consortium, 2024), NPASS (1,150)(Zeng et al., 2018)
- **Molecule** (97): DrugBank (97) (Knox et al., 2024)
- **Activity** (78): Already hydrated (78)
- **Gene** (13,363): NCBI Gene (13,363) (Sayers et al., 2021)
- **Disease** (8,934): OMIM (5,738) (of Genetic Medicine, 2025), SNOMED CT (717) (El-Sappagh et al., 2018), MeSH (385)(Lipscomb, 2000), UMLS (796) (of Medicine , US), Orphanet (1,238)(Weinreich et al., 2008), PharmGKB (59) (Whirl-Carrillo et al., 2021)
- **Phenotype** (6,854): Human Phenotype Ontology (HPO) (6,854) (Talapova et al., 2023)
- **Pathway** (2,128) : Reactome (2,127) (Milacic et al., 2023)
- **Effect** (171): Already Hydrated (171)
- **Side effect** (5,364): Already hydrated (5,364)
- **Indication** (2,080): Already hydrated (2,080)

The distribution of the lengths of the hydrated names can be seen in Figure 9. Most of the classes have relatively normal names besides compound and protein. An example of the largest compound is ‘Amyloid-beta precursor protein (APP) (ABPP) (APPI) (Alzheimer disease...’ which can be seen as a knowledge base issue regarding UniprotKB. Another example is ‘[(2S,3R,4S,5S,6R)-3-[(2S,3R,4...-tetradecahydropicene-4a-carboxylate’, which appears to be a valid compound name from Pubchem Compound.

C Knowledge Graph Reduction

One concern we had with the reduction techniques was changing the distribution of the entity classes, or completely removing entity classes entirely. Because of this we present Figure 10, which shows the distribution of the node types before and after the reduction. This shows, that all classes still remain, with the reduction technique happening at uniform sampling.

D Prompts

In this section, we present the various prompts used throughout this work. As a reminder, we use the Llama-Nemotron-70B model for all LLM-based generation. All of the prompts are available in the

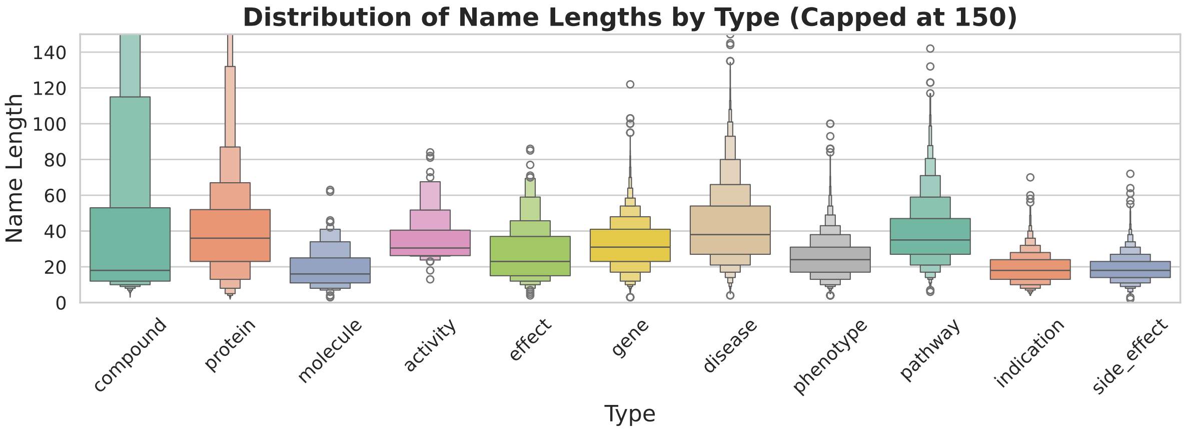


Figure 9: Distribution of hydrated name lengths by node type. 2,539 compound names (7.9%) and 351 protein names (2.4%) exceeded the limit of 150.

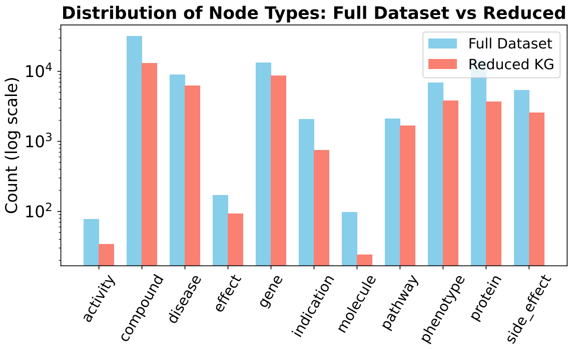


Figure 10: Node type distribution of the OREGANO KG and the reduced version.

GitHub repository, and we highly recommend using the versions provided there. These versions contain additional markup that could not be transferred into this paper. In terms of generation configuration, we utilize LMDeploy, with AWQ quantization, quantization policy of 4, and maximum new tokens of 1,000. Other parameters are default at: *do_sample: False, top_p: 1.0, top_k: 50, min_p: 0.0, temperature: 0.8, repetition_penalty: 1.0,*

D.1 Prompt Based Feature Extraction

The following prompt templates were used to assess the characteristics of QA pairs generated during prompt testing. Each prompt produces a structured JSON response for automated evaluation.

Question Mentions Graphlet Terms

I will give you a question, your job is to tell me if the question mentions any terms that could be related to a graphlet. Please respond only with JSON:
{"justification": justification, "question_mentions_graphlet_terms": True/False}
The justification should be a single string, and "question_mentions_graphlet_terms" must only be a boolean.

Answer Node in Question

I will give you a question and answer pair, your job is to tell me if the answer is present within the question. The answer should have a specific entity not mentioned in the question, if this does not happen return true. Similarly if the answer is a yes/no question or a description about the entities present in the question return true. Please respond only with JSON:
{"justification": justification, "answer_node_in_question": True/False}
The justification should be a single string, and "answer_node_in_question" must only be a boolean.

Answer Mentions Graphlet Terms

I will give you an answer to a question, your job is to tell me if the answer mentions any terms that could be related to a graphlet. Please respond only with JSON:
{"justification": justification, "answer_mentions_graphlet_terms": True/False}
The justification should be a single string, and "answer_mentions_graphlet_terms" must only be a boolean.

Scientifically Accurate Question

I will give you a question, your job is to tell me if the question is scientifically accurate and makes sense from a biological standpoint. The question should sound like an expert is asking it. Further the question should not be trivial. Please respond only with JSON:
{"justification": justification, "scientifically_accurate_question": True/False}
The justification should be a single string, and "scientifically_accurate_question" must only be a boolean.

Answers Question

I will give you a question/answer pair, your job is to tell me if the answer correctly answers the question, and the answer is complete, not lacking in any additional knowledge. Please respond only with JSON:
{"justification": justification, "answers_question": True/False}
The justification should be a single string, and "answers_question" must only be a boolean.

Scientifically Accurate Answer

I will give you an answer to a question, your job is to tell me if the answer is scientifically accurate and makes sense from a biological standpoint. The answer should not be one worded, and be a relatively complete answer, explaining justifications. Please respond only with JSON:
{"justification": justification, "scientifically_accurate_answer": True/False}

The justification should be a single string, and "scientifically_accurate_answer" must only be a boolean.

D.2 Prompt Modules

Below are the various prompt modules used during our ablation test.

Baseline

I will provide you with a graphlet, and your task is to generate a biomedical question-answer pair based on the information within the graphlet.
To enhance the complexity of the question, aim to incorporate as many hops as possible while maintaining coherence.

Simple Example

Example

Question:
What is the primary transmission route for infections like cholera?

Answer:
The fecal-oral route is a primary transmission pathway for infections such as cholera. Contaminated food or water sources facilitate the spread of bacteria like Vibrio cholerae, leading to severe dehydration and gastrointestinal distress.

Instructions Markdown

Instructions:

Analyze Graphlet Instructions

Analyze the graphlet, identifying how nodes are connected and how they might relate in a biomedical context.

Node Types

- Identify three key node types in the graphlet:
1. Question Nodes: These should appear in the question and provide the context for inquiry.
2. Answer Nodes: These should not be explicitly mentioned in the question but must be inferable from the graph structure.
3. Hidden Nodes: These act as logical intermediaries, enabling multi-hop reasoning to reach the answer.

Question Instruction

Construct a question that a biomedical expert might ask, ensuring that:
- The question is phrased naturally as if it were asked in a biomedical research or clinical context. You may mention some graphlet nodes, but do not give away the answer. The question should require multi-step reasoning.
- The answer to the question should be in the graph structure.
- The nodes required to answer the question should not be in the question.
- Ensure scientific relevance, aligning with biomedical terminology and logical reasoning.

Answer Instruction

Your answer should be a scientifically valid response based on the graphlet. Ensure:
- The response is more than a single word; provide a concise yet informative explanation.
- It should justify the answer by connecting relevant biomedical concepts.
- Use precise biomedical terminology while maintaining clarity.

Reflection Instruction

After writing the question and answer, you should reflect on the output and improve the QA pair. If there are no improvements to be made, please repeat the Question/Answer.

Question Evaluation

Question Evaluation Criteria:
- Is the question unambiguous and focused?
- Does the question reflect realistic clinical or research scenarios?
- Does the question require integration of multiple concepts?
- Are terms precise, or could they mislead?
- Is the question too easy?
- Does the question sound natural, or is it too focused on connections from the graph?

Answer Evaluation

Answer Evaluation Criteria:
- Are all facts correct?
- Does the answer address all parts of the question?
- Are key connections explained?
- Does it avoid unsupported claims?
- Are claims supported by pharmacological principles?

Complex Example

Example:

Analysis of Graphlet:

Graphlet contains nodes: [Cholera, Contaminated Water, Fecal-Oral Route, Dehydration]
- Question Node: Cholera
- Hidden Node: Contaminated Water
- Answer Node: Fecal-Oral Route

Initial QA

Question:

What is the primary transmission route for infections like cholera?

Answer:

The fecal-oral route is a primary transmission pathway for infections such as cholera. Contaminated food or water sources facilitate the spread of bacteria like Vibrio cholerae, leading to severe dehydration and gastrointestinal distress.

Reflection

Final QA

Question:

Answer:

Final Analysis

Final Analysis:

Now, analyze the given graphlet and generate a well-formed biomedical question-answer pair.

JSON Format for QA

> Please return the final QA pair in JSON format of:
> { "question": "", "answer": "" }

D.3 Prompt Based Filtering

This prompt is designed to assess the coherence and validity of a QA pair. It evaluates the connections between the entities in the question, generates an answer, and compares it to the previously generated answer. The results are provided in a JSON format, producing two boolean variables to indicate the validity of both the question and the original answer, based on the model's reasoning.

Prompt Filtering

> Evaluate the following question answer pair, first analyze the question, identifying different entities. Then evaluate the various connections between these nodes and identify if the question makes sense from a biomedical standpoint.
>
> After this take the question and try to answer it correctly, being the most scientifically correct.
>
> Finally compare your answer to the answer I provide and tell me if it is scientifically accurate, and completely answers the question.
>
> Present your findings in a JSON string: "{question_reasoning: "", valid_question: true/false, my_answer:"", answer_reasoning:"", original_answer_valid:true/false}"
>
> The fields original_answer_valid and valid_questions must be boolean, the field must be valid JSON, no comments.

E Human evaluation Criteria

In this section we present the exact Likert scale that was used during the human evaluation. The annotation task was explained to the expert evaluator in general terms during a one-to-one meeting, with reference to one of the examples. There were no explicit instructions apart from the Likert scale, since it is self-explanatory.

Scientific Validity of the Question

- **5. Completely valid**: Perfectly aligned with current scientific understanding and uses appropriate terminology.

- **4. Very valid**: Scientifically accurate with only trivial imprecisions.

- **3. Moderately valid**: Contains minor scientific inaccuracies but the core question is scientifically sound.

- **2. Slightly valid**: Major scientific inaccuracies, though some aspects may have scientific merit.

- **1. Not at all valid**: Contains fundamental scientific errors or misconceptions that make the question meaningless or impossible to answer.

Scientific Validity of the Answer

- **5. Completely valid**: Perfectly aligned with current scientific understanding, comprehensive, and appropriately nuanced.

- **4. Very valid**: Scientifically accurate with only trivial imprecisions.

- **3. Moderately valid**: Contains minor scientific inaccuracies but the core information is correct.

- **2. Slightly valid**: Major scientific inaccuracies mixed with some valid information.

- **1. Not at all valid**: Contains fundamental scientific errors, misinformation, or contradicts established knowledge.

Question Complexity

- **5. Very complex**: Requires synthesis of specialized knowledge across multiple biomedical domains or involves cutting-edge research.

- **4. Complex**: Requires advanced knowledge and analysis of biomedical mechanisms or relationships.

- **3. Moderate**: Requires integration of multiple biomedical concepts.

- **2. Simple**: Requires basic understanding of biomedical concepts.

- **1. Very simple**: Basic factual question requiring simple recall of common knowledge.

Specificity of Answer

- **5. Highly specific**: Provides exceptional detail and precision, including quantitative data when appropriate.
- **4. Very specific**: Detailed and precise.
- **3. Appropriately specific**: Right level of detail for the question.
- **2. Somewhat general**: Provides some specifics but lacks precision.
- **1. Too general**: Overly broad and lacks specific details.

Answer Completeness

- **5. Fully complete**: Fully and comprehensively covers every aspect of the question, leaving no gaps.
- **4. Very complete**: Addresses nearly all aspects of the question with appropriate depth and context.
- **3. Moderately complete**: Covers most critical elements but lacks some details or supporting points.
- **2. Partially complete**: Addresses some key elements but omits several important aspects.
- **1. Severely incomplete**: Wrong or addresses only a minimal fraction of what was asked.

Not Qualified to Evaluate

- I cannot evaluate this case with confidence.

F Dataset statistics

Here, we present additional statistics for completeness. Table 1 provides a comprehensive overview of the generation and filtering process for the dataset. Specifically, it details: the total number of each graphlet shape, the downsampling ratio applied to each shape, the resulting number of downsampled graphlets (approximately 10,000), the number of samples generated after z-score filtering, the number of accepted samples after filtering, and the final acceptance ratio.

Regarding the number of questions generated per node, Figure 11 provides a more detailed view, showing the distribution capped at 2,000 questions per node. In total, 41 nodes exceed this cap, with the maximum reaching 9,454 questions. While it is not ideal for certain nodes to appear in such a disproportionately high number of questions, this imbalance stems from the sampling methods used during graphlet selection, which we believe accurately reflect the true underlying distribution.

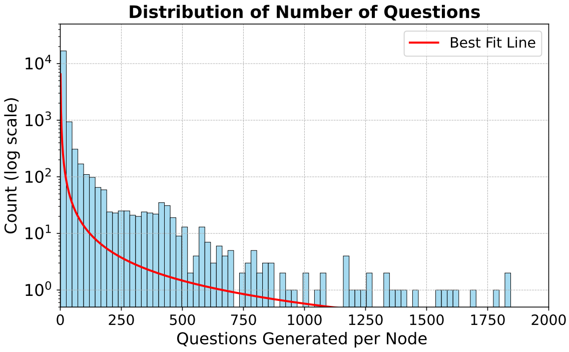


Figure 11: Histogram showing the number of questions generated per mode, limited at 2,000 questions generate per node.

G Human Evaluation Results

The human evaluation results were briefly summarized in the main paper, but here we provide a more detailed analysis. Overall, the evaluations were positive. As previously mentioned, all questions received perfect scores for Scientific Validity, indicating that the questions are meaningful and accurate, even if some are relatively simple.

We do not place strong emphasis on Question Complexity, as it is primarily used to gauge the overall range of difficulty. We are satisfied that the questions span a desirable range—from 3 (Moderate), involving the integration of multiple biomedical concepts, to 4 (Complex), requiring advanced knowledge of biomedical mechanisms or relationships. The presence of some simpler questions is not a concern, as it contributes to the diversity of the dataset and is largely unavoidable given our generation methods. Evaluating question complexity in a more systematic way is left for future work. For now, our focus remains on the scientific quality and correctness of the content.

As mentioned in the paper, 71.7% of samples achieved a minimum score of 3 across the three answer-related criteria, as shown in Figure 12. This increases to 94.4% when considering the average score across the three criteria. However, we believe this average-based metric is not a fully accurate representation—if a sample fails on any one criterion, it should not be considered fully valid.

Table 1: Summary statistics of graphlet generation and filtering. The table reports the total counts of each graphlet shape, downsampling ratios, final counts after downsampling (10,000 per shape), counts after z-score filtering, accepted counts, and acceptance ratios.

ID	Total	Downsampling		Generated	Acceptance	
		Ratio	Count		Total	Ratio
1	2,980,635	3.35×10^{-3}	9,954	9,913	4,544	45.8 %
2	3,702	$1.00 \times 10^{+00}$	3,702	3,690	1,744	47.3 %
3	50,513,861	1.98×10^{-4}	9,826	9,783	4,149	42.4 %
4	41,964,954	2.38×10^{-4}	10,108	10,021	4,475	44.7 %
5	3,609,661	2.77×10^{-3}	10,165	10,103	5,325	52.7 %
6	71,664	1.40×10^{-1}	9,913	9,810	4,485	45.7 %
7	13,537	7.39×10^{-1}	9,939	9,870	4,365	44.2 %
8	11,794	8.48×10^{-1}	10,038	9,948	5,212	52.4 %
9	1,080,297,928	9.26×10^{-6}	9,988	9,817	3,485	35.5 %
10	1,810,874,588	5.52×10^{-6}	9,952	9,806	3,679	37.5 %
11	584,613,716	1.71×10^{-5}	10,126	9,939	4,390	44.2 %
12	922,997	1.08×10^{-2}	10,078	9,874	4,144	42.0 %
13	772,905	1.29×10^{-2}	10,100	9,885	3,897	39.4 %
14	871,384	1.15×10^{-2}	9,946	9,723	4,275	44.0 %
15	46,904	2.13×10^{-1}	10,001	9,823	3,628	36.9 %
16	166,337,860	6.01×10^{-5}	9,946	9,841	5,087	51.7 %
17	239,193	4.18×10^{-2}	10,143	9,949	4,459	44.8 %
18	6,267	$1.00 \times 10^{+00}$	6,267	6,103	2,725	44.7 %
19	225,464	4.44×10^{-2}	10,088	9,878	3,606	36.5 %
20	74,698,349	1.34×10^{-4}	9,894	9,781	5,496	56.2 %
21	55,278	1.81×10^{-1}	9,878	9,629	3,281	34.1 %
22	79,900	1.25×10^{-1}	10,013	9,846	4,533	46.0 %
23	65,548	1.53×10^{-1}	10,031	9,621	4,629	48.1 %
24	11,395	8.78×10^{-1}	9,976	9,741	4,149	42.6 %
25	31,145	3.21×10^{-1}	9,989	9,781	3,759	38.4 %
26	5,617	$1.00 \times 10^{+00}$	5,617	5,292	3,036	57.4 %
27	3,810	$1.00 \times 10^{+00}$	3,810	3,690	1,647	44.6 %
28	18,217	5.49×10^{-1}	10,067	9,577	5,593	58.4 %
29	44,022	2.27×10^{-1}	10,019	9,639	6,059	62.9 %

Therefore, we present the more conservative and fairer minimum-score-based metric as our primary evaluation benchmark. Finally we also show some samples with their corresponding human annotation in Figure 13.

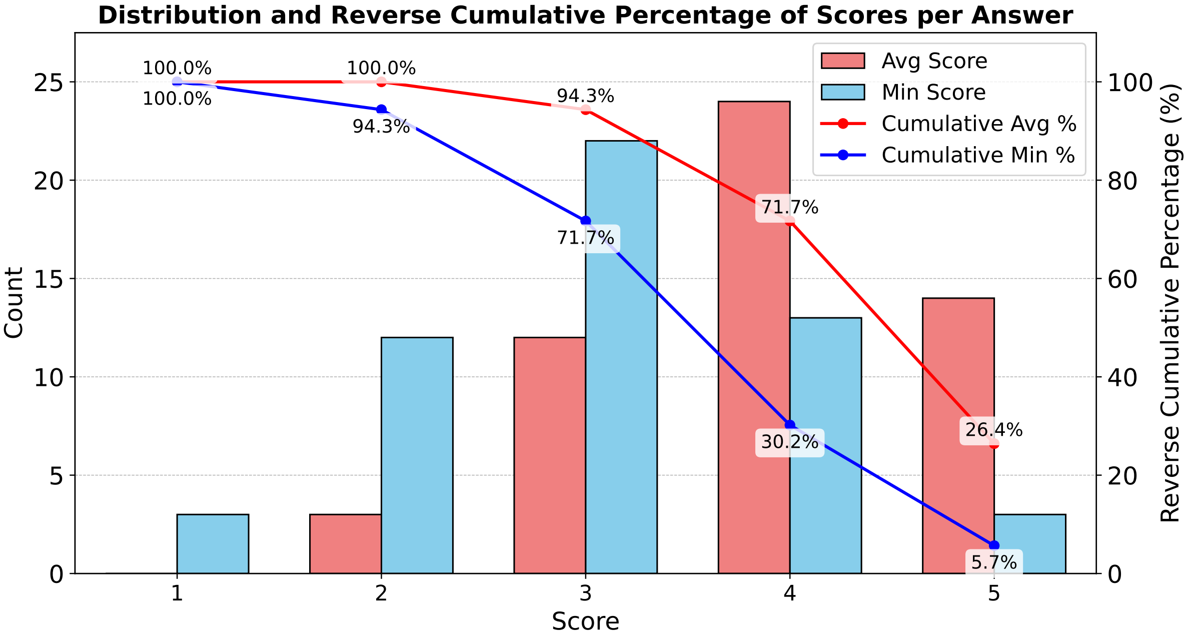
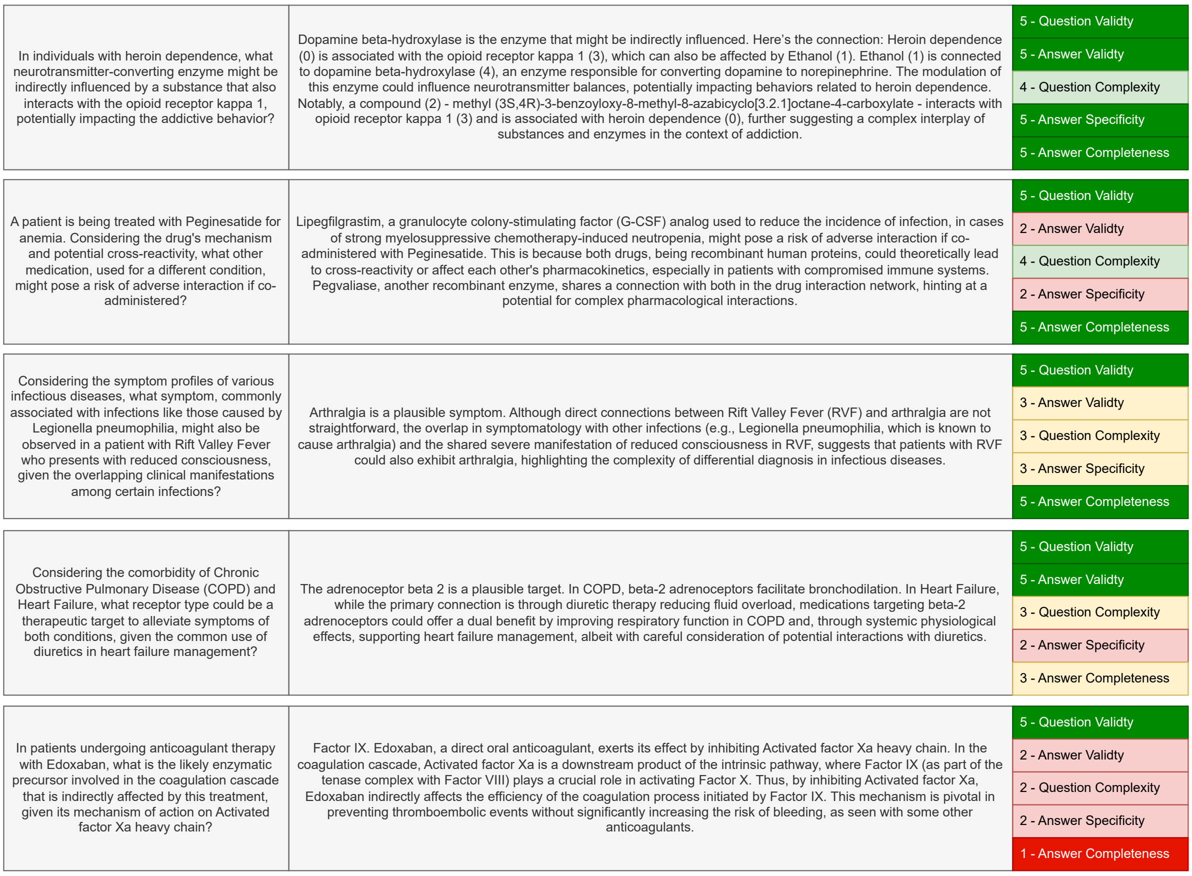


Figure 12: Distribution of minimum and rounded average scores from the human evaluation, based on three answer-related criteria: Answer Completeness, Specificity, and Scientific Validity. The figure also shows the reverse cumulative percentage of scores, indicating the proportion of answers with scores greater than or equal to each bin.



In individuals with heroin dependence, what neurotransmitter-converting enzyme might be indirectly influenced by a substance that also interacts with the opioid receptor kappa 1, potentially impacting the addictive behavior?	Dopamine beta-hydroxylase is the enzyme that might be indirectly influenced. Here's the connection: Heroin dependence (0) is associated with the opioid receptor kappa 1 (3), which can also be affected by Ethanol (1). Ethanol (1) is connected to dopamine beta-hydroxylase (4), an enzyme responsible for converting dopamine to norepinephrine. The modulation of this enzyme could influence neurotransmitter balances, potentially impacting behaviors related to heroin dependence. Notably, a compound (2) - methyl (3S,4R)-3-benzoyloxy-8-methyl-8-azabicyclo[3.2.1]octane-4-carboxylate - interacts with opioid receptor kappa 1 (3) and is associated with heroin dependence (0), further suggesting a complex interplay of substances and enzymes in the context of addiction.	5 - Question Validty 5 - Answer Validty 4 - Question Complexity 5 - Answer Specificity 5 - Answer Completeness
A patient is being treated with Peginesatide for anemia. Considering the drug's mechanism and potential cross-reactivity, what other medication, used for a different condition, might pose a risk of adverse interaction if co-administered?	Lipegfilgrastim, a granulocyte colony-stimulating factor (G-CSF) analog used to reduce the incidence of infection, in cases of strong myelosuppressive chemotherapy-induced neutropenia, might pose a risk of adverse interaction if co-administered with Peginesatide. This is because both drugs, being recombinant human proteins, could theoretically lead to cross-reactivity or affect each other's pharmacokinetics, especially in patients with compromised immune systems. Pegvaliase, another recombinant enzyme, shares a connection with both in the drug interaction network, hinting at a potential for complex pharmacological interactions.	5 - Question Validty 2 - Answer Validty 4 - Question Complexity 2 - Answer Specificity 5 - Answer Completeness
Considering the symptom profiles of various infectious diseases, what symptom, commonly associated with infections like those caused by Legionella pneumophilia, might also be observed in a patient with Rift Valley Fever who presents with reduced consciousness, given the overlapping clinical manifestations among certain infections?	Arthralgia is a plausible symptom. Although direct connections between Rift Valley Fever (RVF) and arthralgia are not straightforward, the overlap in symptomatology with other infections (e.g., Legionella pneumophilia, which is known to cause arthralgia) and the shared severe manifestation of reduced consciousness in RVF, suggests that patients with RVF could also exhibit arthralgia, highlighting the complexity of differential diagnosis in infectious diseases.	5 - Question Validty 3 - Answer Validty 3 - Question Complexity 3 - Answer Specificity 5 - Answer Completeness
Considering the comorbidity of Chronic Obstructive Pulmonary Disease (COPD) and Heart Failure, what receptor type could be a therapeutic target to alleviate symptoms of both conditions, given the common use of diuretics in heart failure management?	The adrenoceptor beta 2 is a plausible target. In COPD, beta-2 adrenoceptors facilitate bronchodilation. In Heart Failure, while the primary connection is through diuretic therapy reducing fluid overload, medications targeting beta-2 adrenoceptors could offer a dual benefit by improving respiratory function in COPD and, through systemic physiological effects, supporting heart failure management, albeit with careful consideration of potential interactions with diuretics.	5 - Question Validty 5 - Answer Validty 3 - Question Complexity 2 - Answer Specificity 3 - Answer Completeness
In patients undergoing anticoagulant therapy with Edoxaban, what is the likely enzymatic precursor involved in the coagulation cascade that is indirectly affected by this treatment, given its mechanism of action on Activated factor Xa heavy chain?	Factor IX. Edoxaban, a direct oral anticoagulant, exerts its effect by inhibiting Activated factor Xa heavy chain. In the coagulation cascade, Activated factor Xa is a downstream product of the intrinsic pathway, where Factor IX (as part of the tenase complex with Factor VIII) plays a crucial role in activating Factor X. Thus, by inhibiting Activated factor Xa, Edoxaban indirectly affects the efficiency of the coagulation process initiated by Factor IX. This mechanism is pivotal in preventing thromboembolic events without significantly increasing the risk of bleeding, as seen with some other anticoagulants.	5 - Question Validty 2 - Answer Validty 2 - Question Complexity 2 - Answer Specificity 1 - Answer Completeness

Figure 13: 5 QA pairs, with their associated human evaluation. Samples selected by hand to show the difference between low scores and high score from the human evaluation.