A SCALABLE LLM FRAMEWORK FOR THERAPEUTIC BIOMARKER DISCOVERY: GROUNDING Q/A GENERATION IN KNOWLEDGE GRAPHS AND LITERATURE

Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Daniel Keitley,

Maria Dermit, Julien Fauqueur*
Digital Science & Innovation, Novo Nordisk
{MBVK,KQTM,LWPH,DKEI,WDRM,JLZF}@novonordisk.com

ABSTRACT

Therapeutic biomarkers are crucial in biomedical research and clinical decision-making, yet the field lacks standardized datasets and evaluation methods for complex, context-dependent questions. To address this, we integrate large language models (LLMs) with knowledge graphs (KGs) to filter PubMed abstracts, summarize biomarker contexts, and generate a high-quality synthetic Q/A dataset. Our approach mirrors biomarker scientists' workflows, decomposing question generation into classification, named entity recognition (NER), and summarization. We release a 24k high quality Q/A dataset and show through ablation studies that incorporating NER and summarization improves performance over using abstracts alone. Evaluating multiple LLMs, we find that while models achieve 96% accuracy on multiple-choice questions, performance drops to 69% on open-ended Q/A, highlighting the need for synthetic data to address the issue of novel discovery. By addressing a critical resource gap, this work provides a scalable tool for biomarker research and demonstrates AI's broader potential in scientific discovery.

1 INTRODUCTION

Scientific discovery in the domain of biomedical research presents a unique set of challenges, particularly when leveraging large language models (LLMs) and autonomous LLM agents Gao et al. (2024). While these models have demonstrated remarkable success in structured fields such as mathematics and code generation, their application in open-ended scientific domains reveals persistent challenges in benchmarking and evaluation Hendrycks et al. (2020). In scientific discovery, reasoning over ambiguous, context-dependent relationships is crucial, yet the ground truth is often neither static nor fully known Park et al. (2023).

Navigating the vast and evolving landscape of scientific literature is pivotal for breakthroughs. However, in open-ended domains, particularly biomedical research, benchmarking and evaluation remain significant challenges Tam et al. (2024). It is through the exploration of current literature and the formulation of precise questions that we begin to unravel complex, not fully understood, scientific concepts.

This challenge is particularly evident in biomedical research, where tasks such as therapeutic biomarker discovery necessitate inferring complex, implicit relationships from heterogeneous literature sources Califf (2018); Polasek & Peck (2024); Kraus (2018). The implicit nature of these relationships—often inferred from subtle contextual cues rather than explicit disease-based mentions—poses a considerable challenge Zhang et al. (2024); Piñero et al. (2015).

The emergence of LLMs presents both opportunities and challenges for biomarker discovery. While LLMs excel in general language understanding, they struggle with specialized tasks such as identi-

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

fying early-stage biomarker signals Liu et al. (2024). To address this, our work introduces a novel framework for generating a large-scale synthetic QA dataset for therapeutic biomarkers. This framework leverages the precision medicine-oriented knowledge graph PrimeKG Chandak et al. (2023) in conjunction with state-of-the-art LLMs, such as GPT 4o-mini. By integrating PrimeKG with LLMs, we aim to identify relevant biomarker contexts using a multi-step prompt engineering approach grounded in graph structures and textual evidence. This approach produces high-quality, domain-specific QA data that surpasses the limitations of purely prompt-driven methods Achiam et al. (2023).

The synthetic dataset serves two key purposes: (1) providing a task-specific benchmark for evaluating LLMs in specialized biomarker reasonin (2) facilitating biomarker discovery by fine tuning LLMs and enabling reasoning over subtle, early-stage associations.

Our contributions:

- **Novel Synthetic Data Pipeline:** We develop an end-to-end pipeline that integrates graph-based retrieval (PrimeKG) with LLM-based summarization to generate a domain-specific QA dataset at scale (Figure 1).
- **First Large-Scale Therapeutic Biomarker QA Dataset:** We publicly release a resource tailored to early-stage biomarker, advancing both synthetic dataset generation and biomedical NLP. It is accessible in Mendeley Data with the identifier: DOI: 10.17632/jmp5n8wnrt.1.
- **Comprehensive Empirical Evaluation:** We provide quantitative and qualitative assessments of the dataset, measuring domain relevance, factual consistency, and diversity while comparing different summarization strategies.
- **Benchmark for LLM Performance:** We evaluate multiple LLMs (OpenAI o3-mini, DeepSeek-R1, OpenAI o1, GPT-4o, Llama 3.3 70b) to highlight strengths and weaknesses in reasoning about complex biomarker relationships.

Ultimately, this work demonstrates that integrating structured knowledge graphs, such as PrimeKG, with LLMs can advance synthetic data generation and task-specific evaluations in biomedicine (seen Figure 2). Our framework in Figure 1 offers a scalable approach to biomarker discovery and beyond.

2 BACKGROUND

2.1 LLM-DRIVEN SYNTHETIC DATA GENERATION

LLMs have become powerful tools for synthetic data generation in natural language processing (NLP) and biomedical research, enabling the creation of scalable, task-specific datasets (Long et al., 2024). Synthetic data generation typically involves three stages: data generation, curation, and evaluation, each ensuring diversity, faithfulness, and relevance to downstream applications.

The generation phase commences with a seed dataset D_{samp} (labeled or unlabeled) and an input prompt $M_p(T, D)$ to guide the pretrained LLM M_p with a task T in generating task-specific data D_{gen}. A well-constructed prompt integrates task definitions (e_{task}), constraints ($e_{\text{condition}}$), and in-context demonstrations (e_{demo}), forming a coherent instruction set (Yu et al., 2024). Iterative refinement of prompts minimizes hallucinations and irrelevant outputs while enhancing diversity (Chung et al., 2023).

The curation phase involves filtering D_{gen} to create a high-quality subset D_{cur}, addressing issues such as noisy labels and redundant samples. Automated filtering methods, including heuristic-based techniques and LLM-based critics, rank sample quality. Additionally, domain experts validate critical subsets of D_{cur}, ensuring scientific and clinical relevance (Seedat et al., 2024).

Finally, the curated dataset is evaluated to quantify its faithfulness, diversity, and task-specific utility. Metrics such as fidelity (alignment with ground truth), diversity (variability across samples), and task performance (usefulness for downstream applications) ensure that the dataset is both reliable and clinically relevant (Chan et al., 2024). In high-stakes domains like biomarker discovery, these metrics are crucial for maintaining scientific rigor and practical applicability.

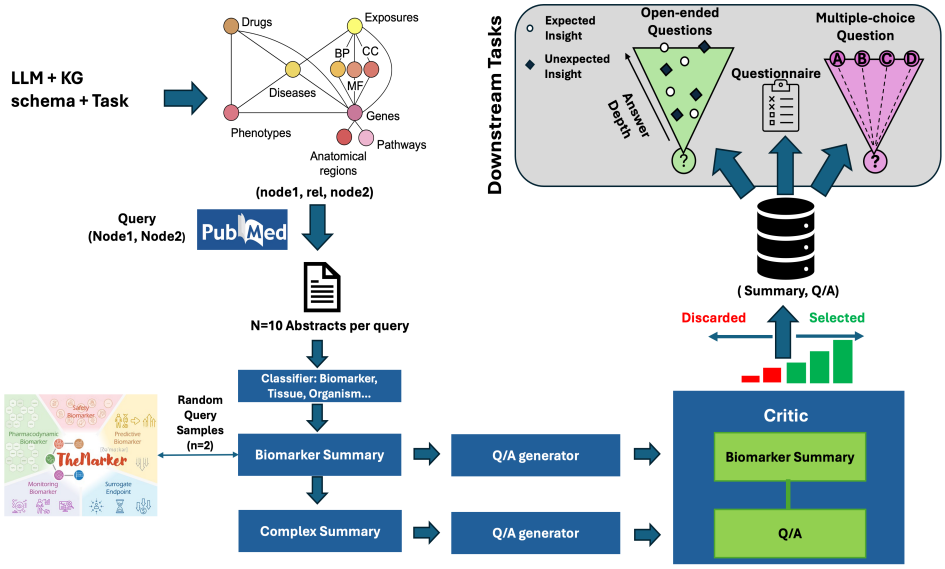

Figure 1: Proposed synthetic question/answer generation framework. Blue boxes represent GPT 4o-mini with specific prompts and a critic LLM, followed by downstream tasks in the gray box. The workflow begins with leveraging the therapeutic biomarker discovery schema to identify relevant relationships, query PubMed abstracts, classify and extract key entities, generate summaries, and filter Q/A pairs. The synthetic questions are used for open-ended and multiple-choice questions, as well as for generating a questionnaire for domain experts.

2.2 Domain Specific Therapeutic Biomarker Dataset

Therapeutic biomarkers are pivotal in drug discovery and clinical applications, offering insights into pharmacological effects, therapy-induced toxicity, and treatment optimization (Kraus, 2018). The Biomarker, Endpoint, and other Surrogate Endpoints (BEST) framework established by the U.S. Food and Drug Administration (FDA) categorizes biomarkers into five key types: diagnostic, predictive, prognostic, pharmacodynamic, and safety markers (Zhang et al., 2024). These categories guide drug development across all stages, from preclinical studies to late-stage clinical trials.

Existing resources, such as *TheMarker*, aggregate biomarker-related information, including a repository of 16,563 biomarkers, predominantly predictive (15,893), with smaller subsets of pharmacodynamic (218) and safety markers (104). However, its reliance on human curation poses critical limitations, including scalability constraints, infrequent updates, and the omission of early-stage biomarkers implicitly described in the literature (Zhang et al., 2024). To overcome these challenges, automated solutions that integrate advanced knowledge representations, such as knowledge graphs, with language models are needed to enable scalable and adaptive biomarker discovery.

3 Methodology

We present a framework (Fig. 1) designed to tackle the challenges associated with identifying therapeutic biomarkers by harnessing synthetic data generation techniques. Biomarkers, often context-dependent and specific to biological conditions, present significant hurdles in discovery due to their implicit relationships in the literature. Our framework integrates KGs, PubMed abstracts, and LLMs, incorporating prompt engineering toolkits Arawjo et al. (2024) and multi-step generation. This integration aims to first identify and process biomarker-related information, then generate summaries, and subsequently create synthetic Q/A datasets from biomarkers. Our approach emphasizes scalability and contextual relevance.

3.1 IDENTIFYING BIOMARKER-RELATED ABSTRACTS

To identify biomarker-related abstracts, we adopted a hybrid approach that combined PrimeKG Chandak et al. (2023), a biomedical KG, with PubMed abstracts and LLMs. PrimeKG encompassed various biological scales, including genes, proteins, diseases, and therapeutic actions, providing a robust foundation for biomarker discovery. Relationships in PrimeKG were represented as edges, $\mathcal{E}_{\text{rel}} = \{(v_i, r_k, v_j)\}$, where v_i and v_j denoted node entities (e.g., genes, diseases), and r_k represented relationship types.

Using the knowledge graph schema and GPTo1, we extracted biomarker-related edges validated by domain experts and generated search queries $Q = \text{GenerateQuery}(v_i, v_j)$ to retrieve abstracts from PubMed. Retrieved abstracts underwent classification using GPT 4o-mini to retain only relevant human therapeutic biomarkers (seen Appendix A.1). The classified abstracts, denoted as $A_{\text{classified}}$, were subjected to Named Entity Recognition (NER) to extract key biological entities, $F_i = \{e_1, e_2, \dots\}$, including tissue types, cell types, mechanisms of action, and co-expressed genes or proteins, which served as features for synthetic summary and Q/A generation.

3.2 GENERATING SUMMARIES AND Q/A PAIRS

Leveraging the classified abstracts and extracted entities, we generated summaries focused on biomarker interactions. Each abstract $a_i \in A_{\text{classified}}$ was summarized with entities F_i guiding the process. To enhance diversity, we sampled two functional descriptions from TheMARker dataset to form D_{samp}, which were utilized in the summarization prompt: $d_{\text{summ}} = M_p(a_i, F_i, D_{\text{samp}})$, where M_p represented the LLM in our case GPT 4o-mini (seen Appendix A.2). In cases where abstracts shared overlapping entities, we synthesized summaries across multiple abstracts to produce richer, multi-contextual representations (seen Appendix A.3). These summaries were then used to generate Q/A pairs (seen Appendix A.4).

3.3 EVALUATION OF SYNTHETIC DATA

Due to the large volume of generated Q/A pairs, direct human evaluation of the entire dataset was infeasible. Summaries were evaluated for faithfulness, completeness, and conciseness, inspired by FineSurE Song et al. (2024), a protocol for assessing sentence-level accuracy and inclusion of key facts (seen Appendix A.5). Summarized text was then used to generate Q/A pairs, ensuring questions focused specifically on biomarker interactions, guided by features F_i, minimizing irrelevant information, and reducing hallucinations, while directing attention to key biomarker relationships. By focusing the evaluation on the interaction guided by F_i, it allowed us to later scrutinize the quality of the Q/A pairs in context.

In the second step of the evaluation, we verified the accuracy and relevance of the generated Q/A pairs using a scoring criteria inspired by DeepSeek-Prover (Xin et al., 2024). Specifically, the GPT 4o-mini was instructed to classify the quality of each Q/A pair into categories such as "poor," "fair," "average quality," "good," and "excellent" (excluding "average quality," "fair," and "poor"), following (seen Appendix A.6). Manual questionnaires of these scores confirmed their alignment with expert intuition.

3.4 EVALUATING LLMS ON BIOMARKER QUESTIONS

We assessed the performance of LLMs in answering biomarker-related questions using a variety of metrics, including multiple-choice and open-ended questions. These questions measured the models' ability to retrieve and reason about biomarker-specific information (Zheng et al., 2023; Kamalloo et al., 2023).

In our evaluation, we employed both multiple-choice and open-ended question formats. For the multiple-choice approach, we created fictitious answers derived from the summaries, while in the open-ended questions, we presented the questions similarly to the way we formulated the real answers, without additional contextual information. This comprehensive evaluation framework aimed to replicate real-world scenarios, providing valuable insights into the LLMs' capacity to handle domain-specific challenges and infer knowledge pertinent to biomarker discovery. The quality of the answer was measured based on biological relevance (seen Appendix A.7).

4 RESULTS

4.1 ABLATION PROMPTING CONTEXT FOR BIOMARKER Q/A GENERATION

In order to understand the components necessary for a successful prompting strategy for Q/A generation, we considered different prompting contexts and evaluated their effect on the quality of synthetic biomarker Q/A generation, focusing on faithfulness (alignment with ground truth) and diversity (variation across generated Q/A pairs), following the methodology of (Liu et al., 2024).

Synthetic biomarker Q/A pairs were generated for a randomly selected subset of 500 relationships sampled from PrimeKG, across six levels of increasing contextual complexity : (1) biomarker generation without any contextual information, (2) inclusion of a single entity representing a potential biomarker, (3) inclusion of two entities describing a biomarker interaction, (4) two entities paired with an abstract where the biomarker interaction is described, (5) multi-step generation using two summaries sampled from *TheMarker* in place of the abstract, and (6) multi-step generation incorporating tissue and cell type classification, mechanisms of action, and co-expressed genes (up-and down-regulated). An example of the different Q/A pair can be seen in section Figure 2.

- Complexity 1:
  - Q: How specific is the biomarker to a particular disease or condition?
  - A: The biomarker exhibits high specificity to the disease, with minimal cross-reactivity observed with other related conditions. Studies indicate that it primarily reflects the disease state and shows consistent results across different populations.
- Complexity 2:
  - Q: How specific is RUNX3 to a particular disease or condition?
  - A: RUNX3 is primarily associated with gastric cancer and has shown significant specificity for this condition, although it may be implicated in other cancers as well. Its expression levels differ in various conditions, which suggests a potential role in differential diagnosis.
- Complexity 3:
  - Q: How specific are RUNX3 and HDAC1 as biomarkers in differentiating between various cancer types?
  - A: RUNX3 and HDAC1 show varying levels of specificity in cancer types, with RUNX3 being more closely associated with gastric cancer, while HDAC1 has implications in a range of cancers including breast and colorectal cancer. Their combined expression patterns can potentially enhance the diagnostic accuracy in particular cancer subtypes.
- Complexity 4:
  - Q: What is the relationship between RUNX3 and HDAC1 in the context of cancer cell proliferation?
  - A: RUNX3 is repressed by histone deacetylation mediated by HDAC1 and hypermethylation of its promoter region. Knockdown of EZH2 reduces trimethylation of histone H3 at lysine 27 and releases HDAC1 from the RUNX3 promoter, leading to an increase in RUNX3 expression and decreased cancer cell proliferation.
- Complexity 5:
  - Q: How does HDAC1 repression of RUNX3 through histone deacetylation and hypermethylation affect the expression of RUNX3 in gastric cancer?
  - A: HDAC1 represses RUNX3, leading to decreased expression of RUNX3 in gastric cancer. This repression is associated with elevated levels of EZH2, which inversely correlates with the expression of RUNX3 and promotes cancer cell proliferation.
- Complexity 6:
  - Q: What role does HDAC1 play in the repression of RUNX3 in gastric cancer cells, and how does this relate to the elevated levels of EZH2 and the mechanisms of histone deacetylation and hypermethylation?
  - A: HDAC1 represses RUNX3 through the mechanisms of histone deacetylation and hypermethylation, resulting in decreased expression of RUNX3. This repression is associated with elevated levels of EZH2, a histone methyltransferase that further inhibits RUNX3 expression. The downregulation of RUNX3 promotes proliferation of gastric cancer cells, highlighting a critical interplay in oncogenic signaling pathways involved in malignancy.

Figure 2: Example of synthetic biomarker Q/A pairs demonstrating generation across six complexity levels, ranging from no contextual information to multi-step generation integrating diverse biological aspects for (RUNX3,HDAC1) entity pair.

In Figure 3, we visualize the effect of these increasing complexity levels, showing two-dimensional t-SNE embeddings for a subset of 100 Q/A pairs. As seen in Figure 3, Level 1 Q/A pairs (in yellow) form a distinct, compact cluster, indicating the limited diversity due to the absence of contextual information. In contrast, Level 6 Q/A pairs exhibit significantly greater dispersion, reflecting the inclusion of multi-step generation, contextual information from *TheMarker*, and detailed classifications. These findings confirm that higher-context prompts facilitate the generation of richer and more diverse Q/A pairs, aligning with expectations for modeling complex biomarker interactions.

Table 1: Q/A generation with increasing levels of complexity, showing the observed fraction of times that a given prompt level was ranked higher than all other levels. Higher complexity levels perform better.

Evaluation of Q/A	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
% of best Q/A	0.0	0.0	0.0	13.5	8.9	77.6

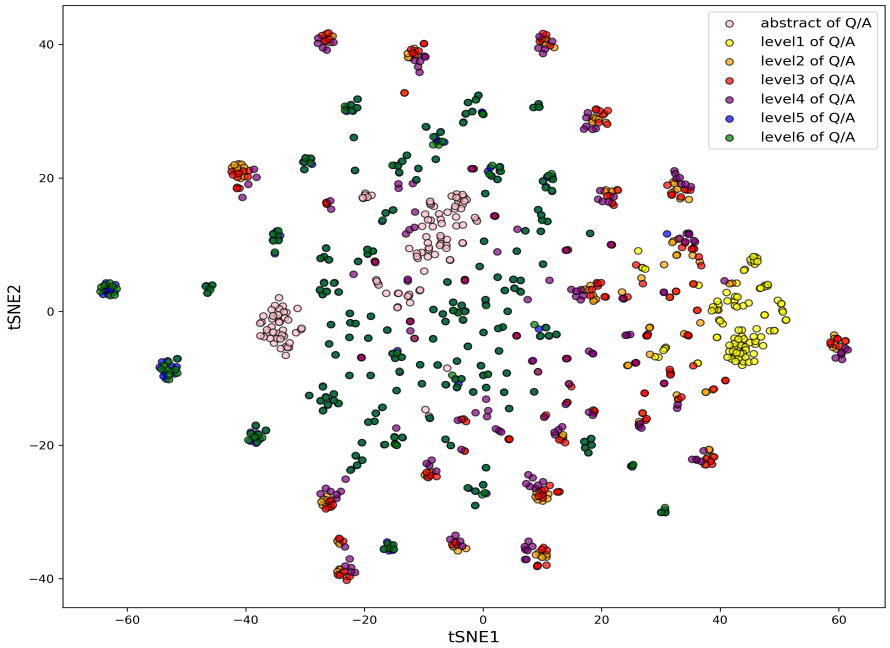

Figure 3: Visualizing the diversity of generated Q/A pairs: t-SNE visualization of a subset of Question/Answer text embeddings, coloured by complexity level (from level 1 in yellow to level 6 in green).

With data generation methods, one of the main factors was the quality of the data generated per attempt and how many attempts it took to get good synthetic data. To test this, we randomly sampled 500 questions and answers equally distributed across the different relationship types that had previously been identified as relevant for identifying useful biomarker questions. Then we let the GPT 4o-mini rank the Q/A pairs from best '1' to worst '6', and Table 1 shows the percentage that each prompting level appeared as the '1', clearly showing the benefit of including the abstract in the prompt from level 4 onwards and the benefit of summarizing the abstracts and extracting useful NER (seen Appendix A.8).

4.2 Evaluation of LLMs on Answering Questions

We wanted to perform an evaluation with two goals: aside from comparing performance of different LLMs, to understand how the difficulty levels differ between our multiple choice and open-ended synthetic Q/A datasets. As indicated in Table 2, the multiple choice questions turned out to be substantially easier to answer by *all* LLMs, with little variation in performance (ranging from 94% to 96%). This is in contrast to open-ended questions, which were significantly more difficult, as aligned with literature (Myrzakhan et al., 2024), allowing us to highlight how the capabilities of various LLMs differ (ranging from the lowest performing Llama 3.3 70B 22% to OpenAI o3-mini 69%). We believe that this highlights the importance of using open-ended scientific questions for both, fine-tuning and evaluating existing LLMs.

Overall, GPT o1, DeepSeek R1 and GPT o3-mini exhibited a notable performance improvement of over 25% compared to other models. The model's ability to break down complex biological concepts into simpler elements was evident in its chain of thought (Guo et al., 2025). Whilst we see

- Question: In what ways does the regulation of LDLR expression by MYLIP influence Lp(a) uptake in human liver cells, particularly in HuH7 cells, and how does this interplay relate to cholesterol metabolism abnormalities observed in elevated Lp(a) level conditions?
- Answer: MYLIP plays a critical role in the regulation of LDLR expression in human liver cells, specifically within the HuH7 cell line. By negatively regulating LDLR, MYLIP decreases the uptake of Lp(a), a lipoprotein associated with cardiovascular risk. This mechanism underscores MYLIP's involvement in cholesterol metabolism, as lower LDLR levels reduce the liver's capacity to clear Lp(a) from circulation. Consequently, elevated levels of Lp(a) may lead to dyslipidemia and increased cardiovascular disease risk, highlighting the importance of MYLIP in lipid homeostasis and potential therapeutic targets in conditions characterized by high Lp(a).
- **OpenAI o3-mini** (Answer): MYLIP (also known as IDOL) is an E3 ubiquitin ligase that actively targets the LDL receptor (LDLR) for degradation, thereby reducing overall LDLR surface expression. Because Lp(a) competes with LDL for LDLR-mediated uptake, reduced LDLR availability lessens the capacity of human liver cells, including HuH7 cells, to internalize Lp(a). In individuals with elevated Lp(a), this downregulation of LDLR by MYLIP may exacerbate cholesterol metabolism imbalances by allowing higher circulating Lp(a) levels. Normally, LDLR activity helps mitigate Lp(a)-associated atherogenic risk by promoting clearance; however, if MYLIP expression is increased or its function is enhanced, LDLR pools become depleted, reducing Lp(a) uptake and contributing to persistent hypercholesterolemia. In HuH7 liver cells, experimental knockdowns or inhibition of MYLIP often restore LDLR presence on the cell surface, resulting in more robust Lp(a) clearance and improved lipid homeostasis. Thus, the MYLIP-LDLR pathway plays a central role in modulating Lp(a) clearance, and its dysregulation directly correlates with the abnormal cholesterol metabolism characteristic of conditions in which Lp(a) levels are elevated.
- **DeepSeek R1** (Answer): The regulation of LDL receptor (LDLR) expression by MYLIP (myosin regulatory light chain-interacting protein, also known as IDOL) in human liver cells, including HuH7 hepatoma cells, directly impacts lipoprotein(a) [Lp(a)] uptake and cholesterol metabolism. MYLIP, an E3 ubiquitin ligase, ubiquitinates LDLR, targeting it for lysosomal degradation, thereby reducing surface LDLR availability. This downregulation diminishes LDLR-mediated endocytosis of LDL and, to a lesser extent, Lp(a), which shares partial LDLR dependency for cellular uptake. In HuH7 cells, MYLIP overexpression reduces LDLR levels, impairing Lp(a) clearance and elevating intracellular cholesterol scarcity, which activates sterol regulatory element-binding protein 2 (SREBP-2) to enhance cholesterol biosynthesis. However, elevated Lp(a) levels common in hypercholesterolemia exacerbate cholesterol metabolism abnormalities by competing with LDL for residual LDLR uptake and promoting pro-atherogenic pathways via oxidized phospholipid-rich Lp(a) particles. This interplay creates a dyslipidemic loop: reduced LDLR activity from MYLIP upregulation impairs Lp(a) and LDL clearance, driving hypercholesterolemia and atherosclerosis risk. Additionally, Lp(a)'s structural similarity to plasminogen may further disrupt hepatic lipid homeostasis through thrombo-inflammatory mechanisms. Thus, MYLIP-mediated LDLR suppression in hepatocytes directly links elevated Lp(a) to cholesterol metabolism dysfunction, highlighting a pathway for therapeutic targeting in cardiovascular diseases.
- **Openai-o1** (Answer): MYLIP, also referred to as IDOL (Inducible Degrader of the LDL Receptor), downregulates LDLR expression by tagging the receptor for ubiquitin-mediated degradation. In human liver cells such as HuH7, this downregulation can significantly reduce the uptake of lipoprotein(a) (Lp(a)) because Lp(a) relies in part on the same receptor pathways as LDL for clearance. With fewer LDLRs available on the cell surface, Lp(a) accumulates in the circulation. Elevated Lp(a) levels exacerbate cholesterol metabolism abnormalities: they can promote an atherogenic environment via both direct lipid accumulation and the heightened inflammatory response triggered by Lp(a). This MYLIP-LDLR regulatory axis thus not only lowers Lp(a) clearance but also contributes to the broader cholesterol imbalance seen in conditions of elevated Lp(a).
- **GPT o4** (Answer): MYLIP, also known as IDOL, is an E3 ubiquitin ligase that regulates the degradation of the low-density lipoprotein receptor (LDLR). In human liver cells, including HuH7 cells, MYLIP can decrease LDLR levels by promoting its ubiquitination and subsequent degradation. This reduction in LDLR expression can lead to decreased uptake of lipoprotein(a) (Lp(a)), as Lp(a) is known to bind to LDLR for cellular uptake. In conditions where Lp(a) levels are elevated, the downregulation of LDLR by MYLIP may exacerbate cholesterol metabolism abnormalities. This is because reduced LDLR levels can lead to decreased clearance of LDL and Lp(a) from the bloodstream, contributing to hypercholesterolemia and associated cardiovascular risks. Therefore, the regulation of LDLR by MYLIP plays a critical role in modulating Lp(a) uptake and influencing cholesterol homeostasis, particularly under conditions of elevated Lp(a) levels.
- **Llama3.3 70b** (Answer):The regulation of LDLR expression by MYLIP influences Lp(a) uptake in human liver cells, particularly in HuH7 cells, by modulating the endocytic pathway. MYLIP, a E3 ubiquitin ligase, targets LDLR for ubiquitination and degradation, thereby reducing LDLR-mediated Lp(a) uptake. In HuH7 cells, which are a model of human liver cells, MYLIP overexpression decreases Lp(a) uptake, while MYLIP knockdown increases it. This interplay is relevant to cholesterol metabolism abnormalities observed in elevated Lp(a) level conditions, as increased Lp(a) levels are associated with reduced LDLR expression and activity. The decreased LDLR expression leads to impaired clearance of Lp(a) from the circulation, contributing to elevated Lp(a) levels. Furthermore, the reduced LDLR activity also impairs cholesterol uptake and metabolism in liver cells, leading to cholesterol accumulation and metabolic abnormalities.

Figure 4: Example of synthetic biomarker pair QA sets, with responses from different LLMs on the same question.

differences across the models specific biology in answers of Figure 4, we believe that the ability to break down complicated biology tasks is what gives these models a notable advantage.

5 DISCUSSION

This paper introduces a novel framework that integrates KGs, LLMs and scientific literature to systematically generate Q/A pairs. By structuring the Q/A process as a multi-step summarization task, our approach produces domain-specific synthetic data that captures key insights more effectively than processing raw abstracts.

Table 2: Evaluation of LLMs as a fraction (%) of correct responses, comparing Open-ended and Multiple Choice question (MCQ) formats across different LLMs. Evaluated on 10k biomarker Q/A pairs.

Question Type	Llama3.3 70B	GPT-4o	OpenAI o1	DeepSeek R1	OpenAI o3-mini
Open-ended	21.7	24.5	63.2	67.7	**69.1**
MCQ	**93.7**	**94.4**	**95.9**	**95.7**	**95.8**

Through embedding analysis, we demonstrate that our method generates Q/A pairs with greater semantic variation, leading to a richer and more diverse representation of knowledge. Additionally, by incorporating LLM-based critics, we enhance question quality, resulting in more informative and challenging benchmarks. We hope that this dataset will serve as a valuable resource for evaluating LLM performance on open-ended biomedical questions, while also revealing limitations in domain-specific reasoning and factual consistency.

A promising avenue for future work is leveraging reinforcement learning (RL) to fine-tune task-specific models on our synthetic dataset before real-world deployment. RL provides a structured learning environment where models can iteratively refine their reasoning strategies by interacting with high-quality synthetic Q/A pairs. The impressive results of DeepSeek R1 and other reasoning models in our paper suggests that the two might be closely linked. This approach offers a way to improve factual consistency and robustness in biomedical reasoning while ensuring that models generalize effectively to unseen scenarios. Moreover, RL training allows exposure to rare and complex biomedical cases that might not be well-represented or explained in curated datasets.

Further validation through expert evaluation will also be critical to assessing the practical relevance of our dataset and model outputs. Collaborating with biomedical researchers to benchmark LLM-generated answers against expert annotations can provide deeper insights into the reliability and real-world applicability of synthetic Q/A data.

REFERENCES

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.

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Robert M Califf. Biomarker definitions and their applications. *Experimental biology and medicine*, 243(3):213–221, 2018.

Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer, and Sam Denton. Balancing cost and effectiveness of synthetic data generation strategies for llms. *arXiv preprint arXiv:2409.19759*, 2024.

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

John Joon Young Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*, 2023.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

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.

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

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*, 2023.

Virginia B Kraus. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nature Reviews Rheumatology*, 14(6):354–362, 2018.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144, 2023.

Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028, 2015.

Thomas M Polasek and Richard W Peck. Beyond population-level targets for drug concentrations: precision dosing needs individual-level targets that include superior biomarkers of drug responses. *Clinical Pharmacology & Therapeutics*, 2024.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. Curated llm: Synergy of llms and data curation for tabular augmentation in low-data regimes. In *Forty-first International Conference on Machine Learning*, 2024.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. Finesure: Fine-grained summarization evaluation using llms. *arXiv preprint arXiv:2407.00908*, 2024.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine*, 7(1):258, 2024.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36, 2024.

Yintao Zhang, Ying Zhou, Yuan Zhou, Xinyuan Yu, Xinyi Shen, Yanfeng Hong, Yuxin Zhang, Shanshan Wang, Minjie Mou, Jinsong Zhang, et al. Themarker: a comprehensive database of therapeutic biomarkers. *Nucleic Acids Research*, 52(D1):D1450–D1464, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

A PROMPTS

This section provides an in depth look at the prompts used in the paper with an explanation, the bold text in the prompts are variables

A.1 CLASSIFYING AND NER OF ABSTRACTS

The primary objective of this task is to initially classify the abstract related to biomarkers within the context of the two entities extracted from the Knowledge Graph. Subsequently, the task involves classifying whether the interaction is associated with human entities. Upon confirming these conditions, the next step is to extract useful NER information to gain a deeper understanding of the interaction.

As a biomarker expert, your task is to classify the interaction between **entity_1** and **entity_2** from the provided abstract into the following biomarker types: **biomarker_types**.
The class of biomarker should be returned as an abbreviation. If there is insufficient evidence to conclude that it's a biomarker in the abstract, please return 'False' as a string.
Classify the organism where the interaction between biomarker **entity_1** and **entity_2** occurs. If there is insufficient evidence or it's not a human organism, please return 'False'.
For cases where the biomarker is classified as true AND the organism is explicitly HUMAN, perform the following classifications:

1. Tissue where the interaction between biomarker **entity_1** and **entity_2** is described (output is a string).
2. Cell type where the interaction between biomarker **entity_1** and **entity_2** is described (output is a string).
3. Mechanism of action of the interaction between biomarker **entity_1** and **entity_2** is described (output is a string).
4. Genes/proteins that are co-expressed in a down-regulated manner with the biomarker (Output is a list - If any information is not contained in the text, please don't return anything for that category).
5. Genes/proteins that are co-expressed in an up-regulated manner with the biomarker (Output is a list - If any information is not contained in the text, please don't return anything for that category).

Ensure that the output summary response is in the demanded JSON format.

A.2 GENERATE SUMMARY

The task is to generate a concise summary based on the two entities, ensuring that the summary remains brief while also considering the inclusion of other potentially relevant entities when available. To ensure that the summary is appropriate, we randomly sample two examples from TheMarker dataset based on the biomarker type.

As a biomarker expert, your task is to summarize the interaction between **entity_1** and **entity_2** in 2-3 lines as they appear in the provided abstracts, considering the following information when available:

1. Biomarker class: **biomarker_class**
2. Organism: **organism**
3. Tissue: **tissue**
4. Cell type: **cell**
5. Mechanism of action of the interaction between biomarker **entity_1** and **entity_2**
6. Genes/proteins that are co-expressed in an up-regulated manner with the biomarker: **gene_protein_up**
7. Genes/proteins that are co-expressed in a down-regulated manner with the biomarker: **gene_protein_down**

The examples of good summaries of biomarkers belonging to the same **biomarker_class** are:

- **theMARKER_example_1**
- **theMARKER_example_2**

Ensure that the output summary response is in the demanded JSON format.

A.3 GENERATE COMPLEX SUMMARY

The task is to generate a combined summary based on multiple abstracts. We ensure that the biology between the abstracts is similar so that it makes sense to combine the information.

As a biomarker expert, your task is to examine the list of abstracts (**abstracts**) to summarize the interaction between **entity_1** and **entity_2** in 2-3 lines as they appear in the provided abstracts, considering the following information when available:

1. Biomarker class: **biomarker_class**
2. Organism: **organism**
3. Tissue: **tissue**
4. Cell type: **cell**
5. Mechanism of action of the interaction between biomarker **entity_1** and **entity_2**
6. Genes/proteins that are co-expressed in an up-regulated manner with the biomarker: **gene_protein_up**
7. Genes/proteins that are co-expressed in a down-regulated manner with the biomarker: **gene_protein_down**

The examples of good summaries of biomarkers belonging to the same **biomarker_class** are:

- **theMARKER_example_1**
- **theMARKER_example_2**

Ensure that the summary is coherent between all abstracts.
If conflicting biology between abstracts is found, return 'False'.
Ensure the output summary response is in the demanded JSON format.

A.4 GENERATE Q/A

The prompt is designed to thoroughly assess the clinical relevance and potential of a therapeutic biomarker of the Q/A in the context of disease diagnosis and treatment. It aims to scrutinize the interdisciplinary potential and innovativeness of the biomarker, ensuring it addresses pressing medical needs and fills identified gaps within the medical and pharmaceutical communities.

You are a biomarker expert, your task is to generate a biomarker question and answer from **x_name**, **y_name** that is CHALLENGING to answer BUT based on evidence from text. MAKE SURE QUESTION AND ANSWER ARE DETAILED
Some extra information that might be available and could be useful when creating the Q/A are the **summary** obtained from abstract for extra context that might be available:

- The classification biomarker: **biomarker_class**
- Organism where biomarker is described: **organism**
- Tissue where biomarker is described: **tissue**
- Cell type where biomarker is described: **cell_type**
- Mechanism of action of biomarker: **mech_of_action**
- Genes/proteins that are co-expressed in a down-regulated manner with a biomarker: **gene_protein_up**
- Genes/proteins that are co-expressed in a down-regulated manner with a biomarker: **gene_protein_down**

MAKE SURE QUESTION is a CHALLENGING QUESTION AND RELATES is about **x_name**, **y_name**. The rest of the information is extra (**organism**, **tissue**, **mech_of_action**, **gene_protein_up**, **gene_protein_down**) and should only be used to make the question harder.
Here are some example questions, use them as reference points but be creative and choose the question and answer that is useful for **x_name**, **y_name**:

- Biomarker Secretion:
  - Is the biomarker actively secreted into the bloodstream, or is its presence in the blood primarily due to passive release from damaged tissues?
  - Could you provide information on the kinetics of biomarker release into the bloodstream under normal and pathological conditions?
- Protein Interactions:
  - Can you provide information about the interaction between protein and protein in the context of biomarker discovery?
  - Are there any known pathways where biomarker and biomarker interact or influence each other's expression?
- Biomarker Specificity:
  - How specific is the biomarker to a particular disease or condition?
  - Are there any cross-reactivities or potential interferences with biomarker in the presence of other related conditions?
- Biomarker Performance:
  - What is the sensitivity and specificity of the biomarker in differentiating between disease stages?
  - Could you provide data on the reproducibility and precision of biomarker in different laboratory settings?
- Biomarker Expression Patterns:
  - Is there evidence of temporal changes in the expression of the biomarker during disease progression?
  - What are the spatial expression patterns of the biomarker in tissue samples from patients with the condition?
- Biomarker Validation:
  - What validation studies have been conducted for the biomarker in independent patient cohorts?
  - Are there any ongoing clinical trials evaluating the utility of the biomarker in predicting treatment response in the condition?

OUTPUT RESPONSE in the demanded JSON format.

A.5 EVALUATE SUMMARY

The task is to evaluate the summaries based on faithfulness, completeness, and conciseness, along with the features F_i.

As a biomarker expert, your task is to assess a biomarker summary (**summary**).
1. Determine if **summary** accurately summarizes the interactions between **entity_1** and **entity_2** as presented in **abstract**. Provide 'True' if accurate, 'False' if not.
Only if 'True', proceed to the following tasks.
2. Evaluate the **summary** of **abstract** based on its faithfulness, completeness, and conciseness.
3. Assess $\mathbf{F_i}$: Identify how many important features are appropriately represented in the **summary**.
Customize your evaluation for the summary, based on the above stringent, and classify it as 'excellent', 'good', 'above average', 'fair', or 'poor'.

A.6 GRADE Q/A

The prompt is designed to thoroughly assess the clinical relevance and potential of a therapeutic biomarker in the Q/A context of disease diagnosis and treatment. It aims to scrutinize the interdisciplinary potential and innovativeness of the biomarker, ensuring it addresses pressing medical needs and fills identified gaps within the medical and pharmaceutical communities.

The prompt is designed to thoroughly assess the clinical and research relevance of a therapeutic biomarker in the Q/A context of disease diagnosis and treatment. It aims to scrutinize the interdisciplinary potential and innovativeness of the biomarker, ensuring it addresses pressing medical needs and fills identified gaps within the medical and pharmaceutical communities.

Evaluate the relevance, novelty, and specificity of the following Q/A pair related to a therapeutic biomarker for drug discovery. Are the biological entities which might be relevant mentioned in the **question** and **answer**?

Question: What is the clinical relevance of this biomarker in the context of disease diagnosis, treatment, or prognosis?
Answer: The biomarker addresses a pressing medical need and is actively researched in the field.

Criteria for Evaluation:
- Clinical Relevance: Does the biomarker address a pressing medical need or concept actively researched in the field of disease diagnosis, treatment, or prognosis?
- Biological Complexity and Predictive Power: Is the biomarker complex enough to provide significant insights into the underlying biological mechanisms of the disease, while also demonstrating strong predictive power for treatment outcomes?
- Interdisciplinary Potential: Does the biomarker offer opportunities for interdisciplinary research, connecting medical science with other fields such as bioinformatics, pharmacology, or genetics?
- Unmet Clinical Needs: Does the biomarker fill an identified gap in disease diagnosis, treatment, or prognosis within the medical and pharmaceutical communities?
- Innovativeness: How innovative is the biomarker? Does it propose new methods, concepts, or applications for disease management or drug development?

Customize your evaluation for each Q/A accordingly, being stringent, and assess it as 'excellent', 'good', 'above average', 'fair', or 'poor'.

A.7 GRADE LLMS RESPONSE TO BIOMAKER QUESTION

Our prompt entails assuming the role of a biomarker examiner tasked with evaluating biomarkers in the early stages for potential drug discovery. The primary objective is to ascertain the correctness of an expert's response to a specific question related to biomarkers. We use the generated response as reference only and evaluate the correct biology strictly to take into consideration the tendency of LLMs hallucinations

You are a biomarker examiner that is examining biomarkers at the early stage for potential drug discovery and you need to check if an expert answered the response correctly. The question is: **question**, and you need to evaluate the answer: **answer**. This is a very challenging task,

and you need to make sure the specific biology is reflected use this as reference **reference answer**. If the correct biology is not reflected, be HARSH. Grade the answer according to 'excellent', 'good', 'above average', 'fair', or 'poor'.

A.8 RANK Q/A OF DIFFERENT PROMPT LEVELS

The prompt is designed to rank a list of Q/A based on the clinical relevance and potential of a therapeutic biomarker in the Q/A context of disease diagnosis and treatment. It aims to assess how incorporating different contexts when generating Q/A affects the quality.

Evaluate the relevance, novelty, and specificity of a list of Q/A pairs related to therapeutic biomarkers for drug discovery. Are the biological entities which might be relevant mentioned in the **question** and **answer**?

Questions and Answers:

1. What is the clinical relevance of this biomarker in the context of disease diagnosis, treatment, or prognosis?

   Answer: The biomarker addresses a pressing medical need and is actively researched in the field.

Criteria for Evaluation:
- Clinical Relevance: Does each biomarker address a pressing medical need or concept actively researched in the field of disease diagnosis, treatment, or prognosis?
- Biological Complexity and Predictive Power: Is each biomarker complex enough to provide significant insights into the underlying biological mechanisms of the disease, while also demonstrating strong predictive power for treatment outcomes?
- Interdisciplinary Potential: Do the biomarkers offer opportunities for interdisciplinary research, connecting medical science with other fields such as bioinformatics, pharmacology, or genetics?
- Unmet Clinical Needs: Do the biomarkers fill identified gaps in disease diagnosis, treatment, or prognosis within the medical and pharmaceutical communities?
- Innovativeness: How innovative are the biomarkers? Do they propose new methods, concepts, or applications for disease management or drug development?

Rank each Q/A pair based on the evaluation, ranging from best '1' (highest rank) to '6' (lowest rank).