

KNOWLEDGE-AUGMENTED LONG-CoT GENERATION FOR COMPLEX BIOMOLECULAR REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Understanding complex biomolecular mechanisms requires multi-step reasoning across molecular interactions, signaling cascades, and metabolic pathways. While large language models (LLMs) show promise in such tasks, their application to biomolecular problems is hindered by logical inconsistencies and the lack of grounding in domain knowledge. Existing approaches often exacerbate these issues: reasoning steps may deviate from biological facts or fail to capture long mechanistic dependencies. To address these challenges, we propose a Knowledge-Augmented Long-CoT Reasoning framework that integrates LLMs with knowledge graph-based multi-hop reasoning chains. The framework constructs mechanistic chains via guided multi-hop traversal and pruning on the knowledge graph; these chains are then incorporated into supervised fine-tuning to improve factual grounding and further refined with reinforcement learning to enhance reasoning reliability and consistency. Furthermore, to overcome the shortcomings of existing benchmarks, which are often restricted in scale and scope and lack annotations for deep reasoning chains, we introduce PrimeKGQA, a comprehensive benchmark for biomolecular question answering. Experimental results on both PrimeKGQA and existing datasets demonstrate that although larger closed-source models still perform well on relatively simple tasks, our method demonstrates clear advantages as reasoning depth increases, achieving state-of-the-art performance on multi-hop tasks that demand traversal of structured biological knowledge. These findings highlight the effectiveness of combining structured knowledge with advanced reasoning strategies for reliable and interpretable biomolecular reasoning.

1 INTRODUCTION

Biological systems are governed by extraordinarily complex mechanisms spanning multiple levels, from protein-protein interactions to intracellular signaling cascades and metabolic pathways (Stoney et al., 2018). Understanding these mechanisms is crucial for areas such as drug discovery, elucidation of disease mechanisms, and analysis of molecular interactions. Crucially, molecular events rarely occur in isolation; instead, they operate within intricate causal chains, making multi-step reasoning indispensable for interpretable mechanistic insights and scientific discovery (Patel et al., 2005; Xu et al., 2024). Meanwhile, the rapid expansion of genomic and proteomic datasets (Ma et al., 2023; Hosseini et al., 2024; Fallahpour et al., 2025), poses unprecedented challenges for computational methods to extract reliable and interpretable reasoning from large-scale biological data.

Recent advances in Large Language Models (LLMs) (Hurst et al., 2024; Guo et al., 2025; Yang et al., 2025) have recently achieved remarkable progress in multi-step reasoning tasks, particularly in domains such as mathematics (Yu et al., 2023), logic (Chen et al., 2024a), and programming (El-Kishky et al., 2025). These advances are largely attributed to Chain-of-Thought (CoT) (Wei et al., 2022; Wang et al., 2022; Kojima et al., 2022; Min et al., 2024) prompting techniques and reinforcement learning strategies, which enable stepwise and logically coherent reasoning. These advances highlight the potential of LLMs to support mechanistic reasoning in biomolecular research, where causal dependencies and multi-hop knowledge traversal are central.

However, direct application of LLMs to biomolecular reasoning remains highly non-trivial (Li et al., 2024; Zhuang et al., 2025): models can produce biologically implausible outputs (Zheng et al., 2024; Fang et al., 2023), generate reasoning chains with logical inconsistencies (Zhang et al., 2024), and

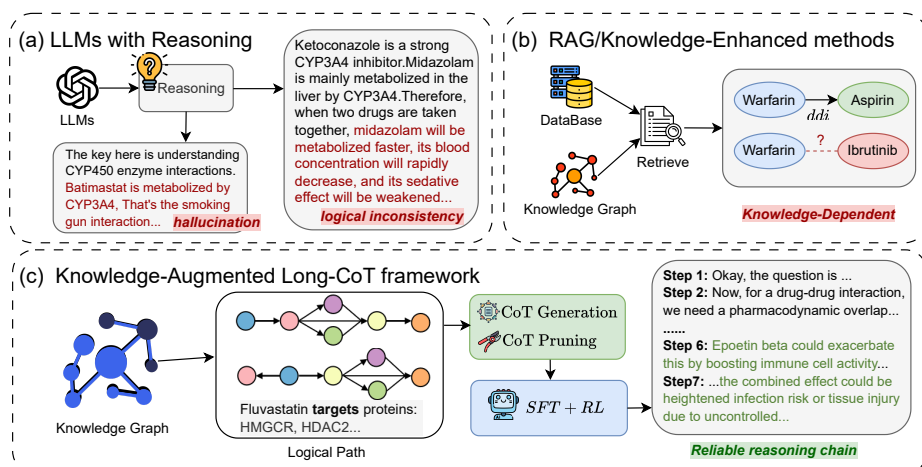


Figure 1: Comparison of different reasoning approaches for biomolecular problems. (a) Reasoning LLMs generate multi-step reasoning chains but often suffer from hallucinations and logical inconsistencies. (b) Retrieval-augmented generation and related knowledge-enhanced methods reduce hallucinations but are knowledge-dependent on the quality and coverage of sources. (c) Our proposed Knowledge-Augmented Long-CoT framework integrates knowledge graph-guided reasoning, enabling logically coherent and reliable reasoning chains for complex biomolecular tasks.

fail to leverage structured knowledge inherent in biological systems (Figure 1a). Although recent knowledge-guided generation methods, such as retrieval-augmented (Soman et al., 2024; Li et al., 2025) or knowledge-graph-based approaches (Liu et al., 2021; Luo et al., 2024; Zhao et al., 2025), partially alleviate hallucination and factual errors, but are inherently knowledge-dependent, relying on the coverage and reliability of external sources (Figure 1b). **Addressing these limitations requires a framework that prioritizes the principled utilization of existing structured knowledge to guide model reasoning, rather than the pursuit of knowledge completeness, thereby enhancing LLM inference and enabling reliable, domain-grounded conclusions.**

To address these challenges, we propose Bio-KCoT, a knowledge-augmented long-CoT reasoning framework for complex biomolecular problems (Figure 1c). At its core, Bio-KCoT introduces a knowledge graph-guided reasoning method that integrates structured biological knowledge to generate biologically plausible and logically coherent reasoning paths. Because raw knowledge graphs contain many redundant or spurious connections, we employ systematic path search and pruning to retain only reasoning chains that are both biologically meaningful and logically complete. These curated chains are then incorporated into supervised fine-tuning to strengthen factual grounding and reinforcement learning with Group Relative Policy Optimization (GRPO) to refine reasoning strategies and enhance robustness on challenging multi-hop tasks. Finally, to enable systematic evaluation, we introduce PrimeKGQA, a benchmark derived from the PrimeKG knowledge graph (Chandak et al., 2023), which spans diverse biomolecular question answering scenarios and task types. PrimeKGQA provides a standardized basis for fair comparison against strong baselines and supports progress in biologically grounded multi-step reasoning. We summarize our contributions as follows:

- We introduce Bio-KCoT, a knowledge-augmented long-CoT reasoning framework for biomolecular problems. It integrates KG-guided path search and pruning to generate coherent reasoning chains used during supervised fine-tuning and reinforcement learning.
- We introduce a high-quality benchmark for biomolecular reasoning, built by systematically collecting and curating knowledge graph-guided reasoning paths that capture multi-level semantic relationships. To ensure fair evaluation, the dataset encompasses diverse biomolecular reasoning scenarios and task types.
- Experimental results on both PrimeKGQA and existing datasets show that while larger closed-source models may retain advantages on simpler tasks, our method achieves substantial gains as reasoning complexity increases, delivering the strongest performance on multi-hop tasks requiring traversal of structured biological knowledge.

2 RELATED WORK

In this section, we provide a comprehensive review of related work, highlighting prior approaches and their relevance to biomolecular reasoning.

2.1 CHAIN-OF-THOUGHT REASONING

Chain-of-Thought reasoning enables LLMs to generate intermediate reasoning steps before arriving at a final answer (Wei et al., 2022; Kojima et al., 2022). This paradigm not only improves the transparency of model predictions, but also enhances performance on tasks requiring multi-step reasoning. Early work introduced few-shot CoT prompting with manually crafted demonstrations (Wei et al., 2022), or simple zero-shot prompting such as “Let’s think step by step” (Kojima et al., 2022). To further strengthen the quality of reasoning, several extensions have been proposed, including self-consistency (Wang et al., 2022), least-to-most prompting (Zhou et al., 2022), complexity-based prompting (Fu et al., 2022), and self-polishing techniques (Xi et al., 2023). More recently, CoT reasoning has been explicitly incorporated into the design of advanced reasoning-oriented LLMs, leading to the emergence of models such as o1 (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), and Google’s Gemini 2.5 Pro (Comanici et al., 2025), which integrate dedicated “thinking” mechanisms and training strategies to enhance multi-step reasoning, coding, and scientific problem solving. Concurrently, the emergence of reasoning-oriented LLMs stimulated interest in distilling their capabilities into smaller student models (<10B), where teacher-generated CoTs are employed as effective supervision signals (Ho et al., 2022; Fu et al., 2023; Magister et al., 2022). However, distilled models still struggle to generalize reasoning beyond benchmark datasets, particularly when factual grounding is critical.

2.2 SCIENTIFIC AND BIOMOLECULAR CoT REASONING

Beyond general NLP benchmarks, the CoT paradigm is increasingly being adapted for complex reasoning in scientific domains. In medicine, models like HuatuoGPT (Chen et al., 2024a) and ReasonMed (Sun et al., 2025) leverage multi-stage, knowledge-infused pipelines to enhance diagnostic reliability (Muennighoff et al., 2025; Dutta & Hsiao, 2024; Zuo et al., 2025). This trend is mirrored in the biomolecular sciences, where models are being developed to reason about intricate biological and chemical processes. These efforts range from incorporating fundamental molecular features like DNA sequences (Fallahpour et al., 2025), protein evolutionary profiles (Liu et al., 2025), and molecular structures (M. Bran et al., 2024; Jang et al., 2024; Narayanan et al., 2025), to modeling higher-level systems such as protein-protein interaction pathways (Jin et al., 2024), tree-structured biological processes (Hsu et al., 2024), and single-cell annotations (Fang et al., 2025).

To facilitate progress, the community relies on benchmarks such as ScienceQA (Lu et al., 2022), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and BioASQ (Krithara et al., 2023). More recently, pathway-centric datasets have been introduced to specifically challenge reasoning over biochemical interactions (Li et al., 2023; Park et al., 2025; Zhao et al., 2025). However, these existing resources often lack the explicit, multi-step rationales tailored to biomolecular reasoning. This scarcity of detailed ground-truth reasoning chains limits their effectiveness in training the very long-CoT models needed to tackle deep biological questions.

2.3 KNOWLEDGE-AUGMENTED REASONING

Although CoT provides a reasoning scaffold, its reliability is fundamentally limited by the internal knowledge of the LLM, which can be prone to hallucination (Huang et al., 2025; Griot et al., 2025; Chen et al., 2024b). To address this limitation, recent work has explored knowledge augmentation, which grounds reasoning in external verified sources such as domain-specific databases or KGs (Chen et al., 2024a; Xie et al., 2024). Among these, KGs are particularly powerful as they provide structured knowledge about entities and their relationships, which can be integrated into LLM reasoning through various techniques, from embedding-based methods to modern hybrid frameworks that enable direct interaction between the LLM and the graph (Guo et al., 2024; Wang et al., 2023; Chandak et al., 2023). This ability to represent complex relational structures makes KGs particularly valuable in domains such as biomedicine, which are defined by intricate networks of interactions.

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

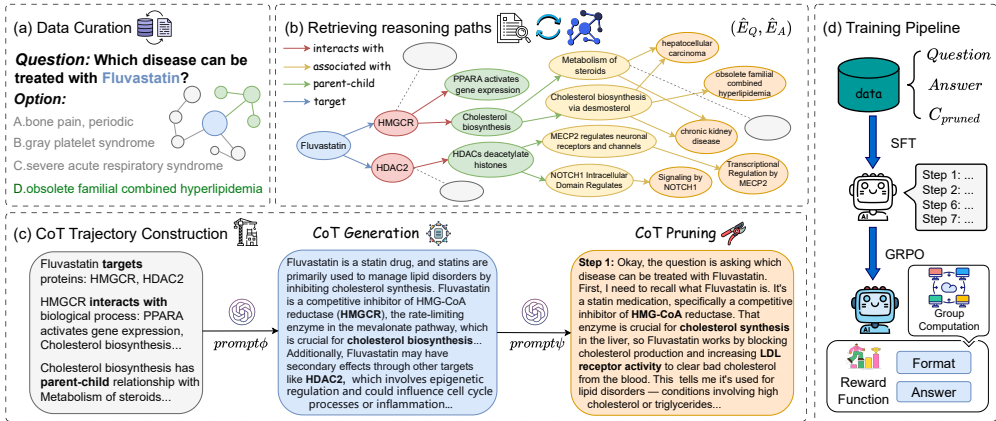


Figure 2: Overview of our proposed framework. (a) **Data Curation**: Given a biomolecular question and candidate answers, we extract entities from the question (E_Q) and the correct answer (E_A). (b) **Retrieving Reasoning Paths**: The extracted entities are mapped onto KG nodes. Reasoning paths $\mathcal{P}(Q, A; d)$ are then retrieved using predefined templates. (c) **CoT Trajectory Construction**: The induced path p provides semantic relations that guide the initial CoT generation. The generated trajectories are further refined through the pruning stage to ensure clarity and accuracy. (d) **Training Pipeline**: The curated (Q, A, C_{pruned}) pairs are used for supervised fine-tuning (SFT), followed by reinforcement learning (GRPO) to align the model’s reasoning and answer generation.

For biomolecular reasoning, where accurate modeling of interactions among genes, proteins, and pathways is essential, knowledge-augmented approaches are particularly promising (Liu et al., 2021; Luo et al., 2024; Zhao et al., 2025). While approaches such as MedReason (Wu et al., 2025) and KG-o1 (Wang et al., 2025) leverage KGs for supervised fine-tuning through shortest-path-based extraction, this strategy is generally well-suited for tasks where relevant evidence can indeed be captured within short paths. However, its effectiveness diminishes for complex problems that require integrating knowledge from disparate parts of the graph. By constraining the evidence to a local neighborhood, it not only risks oversimplifying the reasoning process but also fails to support robust, multi-step inference.

3 METHODOLOGY

The overall workflow, including data curation, reasoning path retrieval, CoT generation, and training pipeline, is illustrated in Figure 2.

3.1 RETRIEVING REASONING PATHS

Given a biomolecular question Q (with problem statement and candidate options) and its corresponding correct answer A , our first step is to retrieve reasoning paths from the pre-constructed biomolecular KG $G = (V, E)$, where V denotes the set of entities and E the relations. The goal is to extract structured reasoning chains \mathcal{P} that bridge the question and the answer.

Entity extraction and mapping. We extract entities mentioned in Q and A , obtaining two sets:

$$E_Q = \{e_i^Q\}_{i=1}^n, \quad E_A = \{e_j^A\}_{j=1}^m. \tag{1}$$

E_Q and E_A denote the sets of entities extracted from the question and the answer, respectively, where n and m are the corresponding numbers of entities. These entities are then mapped to their corresponding nodes in the KG:

$$\hat{E}_Q = \{\hat{e}_i^Q \mid e_i^Q \mapsto \hat{e}_i^Q \in V\}, \quad \hat{E}_A = \{\hat{e}_j^A \mid e_j^A \mapsto \hat{e}_j^A \in V\}. \tag{2}$$

\hat{E}_Q and \hat{E}_A represent the sets of nodes in the KG corresponding to the extracted entities.

Table 1: Illustration of task categories at different difficulty levels. Basic tasks rely on short, direct reasoning in the KG (e.g., drug indication, biological process). Medium tasks involve longer, contextualized reasoning across intermediate entities (e.g., Off-label use, disease–protein associations, side effects). Hard tasks require multi-hop integration of heterogeneous concepts (e.g., contraindications, drug–drug interactions). Here, the blue node denotes the head entity in the question, and the orange node denotes the correct answer entity.

Level	Task Category	Example	Illustration
Basic	Indication	Which disease can be treated with Dalfampridine ?	
	Bioprocess	Which biological process is associated with LIMS1 ?	
Medium	Off-label Use	Which drug is used Off-label for botulism ?	
	Disease-Protein	Which protein is associated with Cutrarino triad ?	
	Side effect	What is a known side effect of Flurbiprofen ?	
Hard	Contraindication	Which disease is contraindication for Nitrogen ?	
	Drug Drug Interaction	Which drug has a drug drug interaction with Piritrexim ?	

Path definition and extraction. Reasoning paths are instantiated from a predefined set of abstract templates, denoted as $\mathbb{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_N\}$. Each template $\mathcal{T}_i \in \mathbb{T}$ encodes a specific topological structure of entity and relation transitions, designed to link a question entity $u \in \hat{E}_Q$ to a candidate answer entity $v \in \hat{E}_A$. These templates accommodate a spectrum of structural complexities, ranging from simple linear chains to intricate non-linear graphs. To illustrate the diversity of these templates, we define three fundamental structural patterns:

- (i) *Linear chain structures*, where entities are connected by a single sequential relation chain:

$$\pi_{\text{lin}}(u, v) = (u = v_0 \xrightarrow{r_1} v_1 \xrightarrow{r_2} v_2 \cdots \xrightarrow{r_d} v_d = v). \quad (3)$$

- (ii) *Divergent structures*, where multiple reasoning trajectories emanate from a common entity:

$$\pi_{\text{div}}(u, v) = \left\{ (u = v_0 \xrightarrow{r_1} v_1), (v_0 \xrightarrow{r_2} v_2 \xrightarrow{r_3} \cdots \xrightarrow{r_d} v_d = v) \right\}. \quad (4)$$

- (iii) *Convergent structures*, where distinct relational branches terminate at a common entity:

$$\pi_{\text{con}}(u, v) = \left\{ (u = v_0 \xrightarrow{r_1} v_1 \xrightarrow{r_2} \cdots v_d = v), (v_0 \xrightarrow{r_3} v_2 \xrightarrow{r_4} \cdots v_d = v) \right\}. \quad (5)$$

In this formulation, d denotes the overall reasoning complexity of a path, which jointly captures (i) the *depth*, corresponding to the maximal length of a relational chain, and (ii) the *breadth*, reflecting the number of admissible branches within the structure.

For a given template $\mathcal{T} \in \{\text{lin}, \text{div}, \text{con}\}$, we define the set of all concrete reasoning paths derived from \mathcal{T} as

$$\text{Inst}(\mathcal{T}; u, v) \triangleq \pi(u, v) \mid \pi(u, v) \models \mathcal{T}. \quad (6)$$

Accordingly, the set of all instantiated paths connecting the question Q and candidate answers A can be formally expressed as

$$\mathcal{P}(Q, A) = \bigcup_{u \in \hat{E}_Q, v \in \hat{E}_A} \left\{ \pi(u, v) \mid \pi(u, v) \in \text{Inst}(\mathcal{T}; u, v), \mathcal{T} \in \mathbb{T}, v_0 = u, v_{|\pi|} = v \right\}, \quad (7)$$

which collects all admissible paths derived from any template in \mathbb{T} connecting entities in Q to candidate answers in A .

Task difficulty characterized by hop depth. We define task difficulty through the complexity metric d , which aggregates two key factors: the maximal depth of the relational chain and the breadth of its structural branches. Greater depths correspond to broader knowledge scopes and higher degrees of informational fragmentation, thereby requiring more demanding semantic integration and logical inference. Accordingly, task difficulty is stratified into discrete levels as a function of d (Eq. 8). Table 1 further illustrates the mapping between representative task categories and their associated depth ranges.

$$\text{Task Difficulty} = \begin{cases} \text{Basic (e.g., Indication, Bioprocess),} & d \leq 5, \\ \text{Medium (e.g., Off-label use, Disease-Protein, Side effect),} & 6 \leq d \leq 7, \\ \text{Hard (e.g., Contraindication, Drug Drug Interaction),} & d \geq 8. \end{cases} \quad (8)$$

In the next stage, these reasoning paths serve as the factual and structural foundation for guiding CoT generation.

3.2 CoT GENERATION AND PRUNING

Building upon the retrieved reasoning paths \mathcal{P} , the next step is to generate a fact-grounded CoT that faithfully reflects the reasoning process underlying the question-answer pair. Each path $\pi \in \mathcal{P}$ provides a structured sequence of entity and relation transitions that can serve as a scaffold for the CoT, ensuring that the reasoning process is grounded in valid biomedical knowledge.

CoT generation. We define a prompt constructor Φ that incorporates the question Q , the correct answer A , the reasoning paths \mathcal{P} . The details of this prompt constructor is illustrated in Figure 6 of Appendix A.3. Given this structured input, the LLM produces a reasoning chain, which we denote as:

$$C = \text{LLM}(\Phi(Q, A, \mathcal{P})). \quad (9)$$

Here, C represents the initial CoT generated by the LLM. The generation process is guided by KG-derived evidence, ensuring that the generated CoT remains aligned with established biomedical knowledge and avoids unsupported associations.

CoT pruning. To enhance the clarity of the factually-grounded CoT(C), which may contain redundant steps or tangential details, we define a pruning prompt constructor Ψ that takes as input the question Q and the initial reasoning chain C . The details of Ψ is illustrated in Figure 7 of Appendix A.3. This prompt instructs the LLM to refine C by removing unnecessary reasoning steps and preserving only the essential logical chain leading to the answer:

$$C_{\text{pruned}} = \text{LLM}(\Psi(Q, C)). \quad (10)$$

Here, C_{pruned} represents the refined reasoning chain that retains only the essential logical steps. Through this prompt-based pruning strategy, the reasoning process becomes more concise and interpretable while maintaining accuracy. A concrete example is provided in Figure 5 of Appendix A.3.

In summary, by leveraging KG-derived paths \mathcal{P} , this two-stage process generates a concise and evidence-grounded reasoning chain C_{pruned} , achieving both factual correctness and interpretability.

3.3 SUPERVISED FINE-TUNING AND REINFORCEMENT LEARNING

After building the reasoning dataset, we conduct supervised finetuning and reinforcement learning. Through the SFT stage, the base model (Qwen3-4B and Qwen3-8B) are trained to follow KG-guided templates, generate coherent multi-step reasoning, and ensure output validity for complex biomolecular questions. This stage effectively mitigates reward sparsity, establishing a robust foundation for the model to efficiently refine its reasoning capabilities during the reinforcement learning phase.

For the reinforcement learning phase, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which leverages intra-group advantage estimation to eliminate the need for a separate critic network. The training process for each input x involves sampling G candidate responses from the current policy and optimizing it using a clipped objective function with group-normalized advantages (the detailed formulation is provided in Appendix A.1).

To simultaneously promote structured reasoning and factual accuracy, we design a composite reward function that enforces both the output format and the correctness. Each response is required to follow a reasoning template, containing intermediate reasoning in `<think>...</think>` and the final answer in `<answer>...</answer>`. The reward is composed of two parts:

$$R_{\text{format}} = \begin{cases} 1, & \text{if format is valid,} \\ 0, & \text{otherwise,} \end{cases} \quad R_{\text{answer}} = \begin{cases} 5, & \text{if the predicted answer is correct,} \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

and the total reward is $R_{\text{reward}} = R_{\text{format}} + R_{\text{answer}}$. This reward structure ensures that answer correctness grants a primary reward of 5, while format adherence provides an additional reward of 1. In this way, these incentives promote both syntactic validity and semantic accuracy, guiding the model toward reliable reasoning and outputs.

4 EXPERIMENT AND RESULTS

4.1 EXPERIMENTAL SETUP

Dataset and Benchmark Existing biomedical QA benchmarks (e.g., BioASQ(Krithara et al., 2023), BiomixQA(Soman et al., 2024)) remain limited in scope: they are relatively small and lack explicit annotations for multi-hop reasoning over structured knowledge. These shortcomings restrict their ability to evaluate the deep reasoning capacities of the models. In contrast, we introduce PrimeKGQA, a benchmark constructed from PrimeKG (Chandak et al., 2023). QA pairs are generated using a template-based approach: Each question is formed from a head–relation pair, with the corresponding tail entity serving as the answer (Figure 2a). The tasks cover diverse biomedical categories(e.g., diseases, drugs, genes, pathways) and exhibit varying levels of reasoning difficulty. In particular, the path length in PrimeKG naturally reflects the complexity of the reasoning required to answer a question (Table 1). To prevent data leakage, the dataset is partitioned by head entities, ensuring no overlap between training and test sets. In total, PrimeKGQA consists of 6,710 QA pairs, with 3,500 for supervised fine-tuning, 1,500 for reinforcement learning, and 1,710 for evaluation. Detailed statistics of the test data are provided in the Appendix A.2. The QA pairs are in the form of single-choice questions with four options, and the evaluation metric is accuracy.

Baselines For the empirical validation of our proposed data and training methodology, we conduct a comparative analysis against a comprehensive set of baselines. To enable a structured comparison, these baselines are grouped into three categories: (1) Open-source model Qwen3-4B and Qwen3-8B (Yang et al., 2025) (base model baseline); (2) Closed-source models, including GPT-4o-mini, GPT-4o (Hurst et al., 2024), and Gemini 2.5 Pro (Comanici et al., 2025); (3) Reasoning-oriented models, such as o1 (Jaech et al., 2024), o3-mini, and Deepseek-R1 (Guo et al., 2025).

Training and Inference We perform further training on the Qwen3-4B and Qwen3-8B models (Yang et al., 2025). The training process follows a two-stage paradigm, consisting of full finetuning and reinforcement learning(RL), using the PrimeKGQA dataset (see Section 4.1). In the finetuning stage, the model is trained for 4 epochs with a learning rate of 1.0×10^{-5} , a per-device batch size of 2, and a maximum input length of 4096 tokens. For the subsequent RL phase, we transition to a parameter-efficient approach, employing Low-Rank Adaptation (LoRA) with a rank of 32. The model undergoes one epoch of training with a learning rate of 1.0×10^{-6} , a per-device batch

Table 2: Performance comparison of Open-source, Closed-source, Reasoning-oriented models, and our proposed model on the PrimeKGQA benchmark test set. The table shows accuracy scores across task categories of varying difficulty levels. Higher values indicate better performance.

Type	Model	Basic			Medium				Hard			All Avg.
		Indication	Bio-process	Avg.	Off-label use	Disease-Protein	Side effect	Avg.	Contraindication	DDI	Avg.	
Open Source	Qwen3-4B	0.768	0.523	0.606	0.710	0.573	0.616	0.618	0.438	0.503	0.473	0.568
	Qwen3-8B	0.845	0.653	0.749	0.671	0.600	0.516	0.586	0.452	0.537	0.498	0.601
Closed Source	GPT-4o-mini	0.903	0.740	0.796	<u>0.903</u>	0.713	0.844	0.801	0.564	<u>0.683</u>	0.625	0.743
	Gemini 2.5 Pro	<u>0.910</u>	0.850	0.880	0.710	0.803	0.804	0.783	0.664	0.647	0.655	0.768
	GPT-4o	0.890	0.873	0.879	0.897	0.807	0.844	<u>0.840</u>	0.676	0.623	0.650	0.789
Reasoning oriented	Deepseek-R1	0.897	0.777	0.837	0.839	0.750	0.803	0.788	0.532	0.625	0.583	0.735
	o3-mini	0.897	<u>0.877</u>	<u>0.884</u>	0.852	0.857	0.772	0.826	0.580	0.663	0.625	0.777
	o1	0.936	0.893	0.908	0.832	<u>0.840</u>	0.844	<u>0.840</u>	0.696	0.703	<u>0.700</u>	<u>0.813</u>
Ours	Bio-KCoT(4B)	0.897	0.773	0.815	0.890	0.743	0.912	0.835	<u>0.804</u>	0.610	0.698	0.786
	Bio-KCoT(8B)	0.901	0.807	0.839	0.930	0.793	<u>0.908</u>	0.864	0.848	0.643	0.736	0.816

size of 4, and a maximum input length of 4096 tokens. During inference, all models are prompted to generate CoT style responses.

4.2 MAIN RESULTS

Table 2 reports the experimental results on the PrimeKGQA benchmark. On basic-level tasks, advanced closed-source LLMs still maintain a leading position, primarily due to their extensive parameter knowledge and strong factual memory capabilities. However, as task difficulty increases, this advantage gradually diminishes and is eventually surpassed. On medium-level tasks, the Bio-KCoT 8B model achieves an average score of 0.864, outperforming both o1 and GPT-4o, which each scored 0.840. Specifically, in the Off-label use subtask, Bio-KCoT 8B achieves 0.930, significantly higher than GPT-4o’s 0.897. In the side effect subtask, the 4B and 8B versions of Bio-KCoT achieve 0.912 and 0.908, respectively, far exceeding the 0.844 scores of GPT-4o and o1.

On the most complex hard-level tasks, Bio-KCoT 8B continues to lead with an overall score of 0.736. Its advantage is more pronounced in the contraindication subtask, where it achieves 0.848, substantially ahead of o1 at 0.696. The experiments also show that Bio-KCoT’s performance improves with increased model size, demonstrating a favorable scale-up property. These results indicate that, with careful framework design and the incorporation of domain knowledge, even models with relatively small parameter counts can achieve performance on complex biomedical tasks that matches or exceeds that of some closed-source large models.

In addition, we observe that the performance of most baseline methods declines progressively from the basic to the medium and hard levels as task difficulty increases. In contrast, our method performs better on medium tasks than on basic ones, and its results on hard tasks remain comparable to those on the basic level. This suggests that the incorporation of long chains of thought becomes particularly advantageous as task complexity increases, enabling the model to leverage reasoning chains more effectively when shallow memorization is insufficient. By generating knowledge-guided reasoning chains, the model can gradually unfold intermediate steps and integrate structured knowledge with semantic information, thereby addressing high-complexity problems more effectively. To identify limitations for future improvement, see the error analysis in the Appendix A.5.

4.3 GENERALIZATION TO OTHER BENCHMARKS

To systematically evaluate the cross-task generalization ability of our method, we further conduct experiments on three external biomolecular question-answering datasets (results shown in Figure 3a). We directly evaluate the model on these out-of-distribution datasets without any further training. These benchmarks span different dimensions of biomolecular challenges: BioASQ (Krithara et al., 2023) reflects real expert information needs based on MEDLINE; BiomixQA (Soman et al., 2024) consists of multiple-choice questions derived from SPOKE disease–gene associations to assess com-

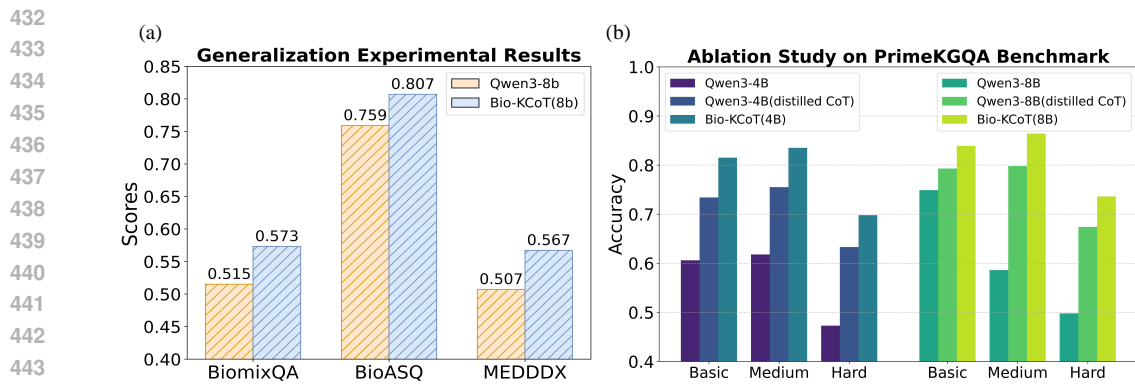


Figure 3: (a) Generalization results on additional biomedical benchmarks(BiomixQA, BioASQ, and MEDDDX). (b) Ablation study on the PrimeKGQA benchmark against the distilled CoT baseline.

plex relational reasoning; and MEDDDX (Su et al., 2024) leverages STaRK-Prime (Wu et al., 2024) to construct closely related distractors for evaluating fine-grained semantic discrimination.

Across these mutually distinct tasks, Bio-KCoT(8B) consistently achieves significant improvements over the base model Qwen3-8B. Notably, Bio-KCoT requires neither task-specific training nor complete external retrieval. Instead, our approach leverages knowledge graph-guided long chains of thought to effectively activate the model’s intrinsic reasoning capacity. This mechanism allows the model to derive reliable answers without depending on fully comprehensive external knowledge, while stably transferring its capabilities to new tasks and domains. These results demonstrate that long CoT not only enhances reasoning performance on in-domain tasks but also exhibits stronger cross-task generalization, laying the foundation for deployment in complex real-world scenarios.

4.4 ABLATION STUDY

We conduct ablation studies to analyze the effectiveness of the proposed knowledge-guided long chain-of-thought generation method. Specifically, we compare the following approaches: (1) the base model Qwen3, on which all our methods are built; (2) Qwen3 (distilled CoT), where the supervised fine-tuning stage employs simple reasoning chains distilled from LLMs rather than knowledge-guided ones; and (3) our proposed Bio-KCoT method. All three approaches are evaluated with models of 4B and 8B parameters.

The experimental results are shown in Figure 3b. Analysis of the experimental results reveals a distinct performance hierarchy that validates our proposed approach. While applying a two-stage training paradigm (SFT and RL) with distilled CoT data yields a significant performance uplift compared to the original Qwen3 base model, our model, trained with Bio-KCoT, consistently and substantially outperforms this distilled CoT counterpart across all task difficulties. Crucially, the superior performance of our model validates the efficacy of our proposed KG-guided reasoning framework, demonstrating its advanced capability in navigating complex, multi-step problems. See the Appendix A.4 for detailed results.

5 CONCLUSION

We introduce Bio-KCoT, a knowledge-augmented long chain-of-thought reasoning framework for complex biomolecular problems. Our approach tackles the issues of unreliable knowledge and logical errors in large language models by using structured knowledge to guide reasoning, without depending on complete external information. Specifically, Bio-KCoT employs a knowledge graph-guided path search and pruning strategy to construct biologically meaningful and logically coherent reasoning chains. Further, we introduce PrimeKGQA, a benchmark dataset covering diverse biomolecular task types with various difficulties. Experimental results demonstrate that Bio-KCoT improves performance on this benchmark, particularly on hard-level questions, and maintains strong generalization on out-of-distribution datasets without additional training.

REFERENCES

- 486
487
488 Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision
489 medicine. *Scientific Data*, 10(1):67, 2023.
- 490
491 Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou,
492 and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint*
493 *arXiv:2412.18925*, 2024a.
- 494
495 Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou
496 Wang. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint*
497 *arXiv:2407.13301*, 2024b.
- 498
499 Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit
500 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
501 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
502 bilities. *arXiv preprint arXiv:2507.06261*, 2025.
- 503
504 Abhishek Dutta and Yen-Che Hsiao. Adaptive reasoning and acting in medical language agents.
505 *arXiv preprint arXiv:2410.10020*, 2024.
- 506
507 Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan,
508 Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. Competitive programming
509 with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.
- 510
511 Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Hao-
512 nan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing
513 multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.
- 514
515 Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. Domain-
516 agnostic molecular generation with chemical feedback. *arXiv preprint arXiv:2301.11259*, 2023.
- 517
518 Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Aidong Zhang,
519 Jiawei Han, and Zhiyong Lu. Cell-o1: Training llms to solve single-cell reasoning puzzles with
520 reinforcement learning. *arXiv preprint arXiv:2506.02911*, 2025.
- 521
522 Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting
523 for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.
- 524
525 Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language
526 models towards multi-step reasoning. In *International Conference on Machine Learning*, pp.
527 10421–10430. PMLR, 2023.
- 528
529 Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models
530 lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642,
531 2025.
- 532
533 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
534 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
535 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 536
537 Lingbing Guo, Zhongpu Bo, Zhuo Chen, Yichi Zhang, Jiaoyan Chen, Lan Yarong, Mengshu Sun,
538 Zhiqiang Zhang, Yangyifei Luo, Qian Li, et al. Mkg1: Mastery of a three-word language. *Ad-
539 vances in Neural Information Processing Systems*, 37:140509–140534, 2024.
- 533
534 Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers.
535 *arXiv preprint arXiv:2212.10071*, 2022.
- 536
537 Ramtin Hosseini, Siyang Zhang, and Pengtao Xie. Text2protein: A generative model for designated
538 protein design on given description. 2024.
- 539
540 Chi-Yang Hsu, Kyle Cox, Jiawei Xu, Zhen Tan, Tianhua Zhai, Mengzhou Hu, Dexter Pratt, Tianlong
541 Chen, Ziniu Hu, and Ying Ding. Thought graph: Generating thought process for biological
542 reasoning. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 537–540, 2024.

- 540 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
541 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
542 models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information*
543 *Systems*, 43(2):1–55, 2025.
- 544 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
545 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
546 *arXiv:2410.21276*, 2024.
- 547 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
548 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv*
549 *preprint arXiv:2412.16720*, 2024.
- 550 Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. Chain-of-thoughts for molecular understanding.
551 *arXiv preprint arXiv:2410.05610*, 2024.
- 552 Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What dis-
553 ease does this patient have? a large-scale open domain question answering dataset from medical
554 exams. *Applied Sciences*, 11(14):6421, 2021.
- 555 Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan
556 Du, and Yongfeng Zhang. Prollm: protein chain-of-thoughts enhanced llm for protein-protein
557 interaction prediction. *bioRxiv*, pp. 2024–04, 2024.
- 558 Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A
559 dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- 560 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
561 language models are zero-shot reasoners. *Advances in neural information processing systems*,
562 35:22199–22213, 2022.
- 563 Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras.
564 Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10
565 (1):170, 2023.
- 566 Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. Biomedrag: A retrieval augmented large
567 language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769, 2025.
- 568 Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang,
569 Le Song, and Yu Li. Progress and opportunities of foundation models in bioinformatics. *Briefings*
570 *in Bioinformatics*, 25(6):bbae548, 2024.
- 571 Yanjing Li, Hannan Xu, Haiteng Zhao, Hongyu Guo, and Shengchao Liu. Chatpathway: Conver-
572 sational large language models for biology pathway detection. In *NeurIPS 2023 AI for science*
573 *Workshop*, 2023.
- 574 Chaoyong Liu, Linlin Chao, Shaomin Ji, Hao Wang, Taorui Jiang, Zhangyang Gao, Yucheng Guo,
575 Ming Yang, and Xiaoming Zhang. Proteinreasoner: A multi-modal protein language model with
576 chain-of-thought reasoning for efficient protein design. *bioRxiv*, pp. 2025–07, 2025.
- 577 Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Martin Ringsquandl, Rime Raissouni, and Volker
578 Tresp. Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In *Euro-*
579 *pean Semantic Web Conference*, pp. 375–391. Springer, 2021.
- 580 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafford,
581 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for
582 science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521,
583 2022.
- 584 Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. Graph-
585 constrained reasoning: Faithful reasoning on knowledge graphs with large language models.
586 *arXiv preprint arXiv:2410.13080*, 2024.

- 594 Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe
595 Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelli-*
596 *gence*, 6(5):525–535, 2024.
- 597
- 598 Chang Ma, Haiteng Zhao, Lin Zheng, Jiayi Xin, Qintong Li, Lijun Wu, Zhihong Deng, Yang Lu,
599 Qi Liu, and Lingpeng Kong. Retrieved sequence augmentation for protein representation learning.
600 *bioRxiv*, pp. 2023–02, 2023.
- 601
- 602 Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn.
603 Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.
- 604
- 605 Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng
606 Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction
607 report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.
- 608
- 609 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
610 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
611 scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- 612
- 613 Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte,
614 Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodrigues, and Andrew D White. Training
615 a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*, 2025.
- 616
- 617 OpenAI. Learning to reason with llms, 2024. URL [https://openai.com/index/
618 learning-to-reason-with-llms](https://openai.com/index/learning-to-reason-with-llms).
- 619
- 620 Gilchan Park, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Shinjae Yoo, and Shantenu
621 Jha. Comparative performance evaluation of large language models for extracting molecular
622 interactions and pathway knowledge. *Journal of Computational Biology*, 2025.
- 623
- 624 Vimla L Patel, José F Arocha, and Jiajie Zhang. Thinking and reasoning in medicine. *The Cam-*
625 *bridge handbook of thinking and reasoning*, 14:727–750, 2005.
- 626
- 627 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
628 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
629 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 630
- 631 Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom,
632 Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomed-
633 ical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40
634 (9):btae560, 2024.
- 635
- 636 Ruth Stoney, David L Robertson, Goran Nenadic, and Jean-Marc Schwartz. Mapping biological
637 process relationships and disease perturbations within a pathway network. *NPJ systems biology
638 and applications*, 4(1):22, 2018.
- 639
- 640 Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clev-
641 ert, and Marinka Zitnik. Knowledge graph based agent for complex, knowledge-intensive qa in
642 medicine. *arXiv preprint arXiv:2410.04660*, 2024.
- 643
- 644 Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing
645 Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for
646 advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025.
- 647
- 648 Nan Wang, Yongqi Fan, Zongyu Wang, Xuezhi Cao, Xinyan He, Haiyun Jiang, Tong Ruan, Jing-
649 ping Liu, et al. Kg-o1: Enhancing multi-hop question answering in large language models via
650 knowledge graph integration. *arXiv preprint arXiv:2508.15790*, 2025.
- 651
- 652 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
653 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
654 *arXiv preprint arXiv:2203.11171*, 2022.

- 648 Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala,
649 and Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graphs.
650 *arXiv preprint arXiv:2310.03320*, 2023.
- 651
- 652 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
653 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
654 *neural information processing systems*, 35:24824–24837, 2022.
- 655 Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu,
656 Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning
657 steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.
- 658 Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis Ioan-
659 nidis, Karthik Subbian, James Y Zou, and Jure Leskovec. Stark: Benchmarking llm retrieval on
660 textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37:
661 127129–127153, 2024.
- 662
- 663 Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Tao Gui, Qi Zhang, and Xuanjing
664 Huang. Self-polish: Enhance reasoning in large language models via problem refinement. *arXiv*
665 *preprint arXiv:2305.14497*, 2023.
- 666 Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing,
667 James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigran-
668 ular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.
- 669
- 670 Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li,
671 Hanqi Jiang, Yi Pan, Junhao Chen, et al. Towards next-generation medical agent: How o1 is
672 reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*, 2024.
- 673 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,
674 Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint*
675 *arXiv:2505.09388*, 2025.
- 676
- 677 Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhen-
678 guo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions
679 for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- 680 Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao
681 Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological
682 & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024.
- 683
- 684 Haiteng Zhao, Chang Ma, Fangzhi Xu, Lingpeng Kong, and Zhi-Hong Deng. Biomaze: Bench-
685 marking and enhancing large language models for biological pathway reasoning. *arXiv preprint*
686 *arXiv:2502.16660*, 2025.
- 687 Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T May, Geoffrey I Webb, Shirui Pan,
688 and George Church. Large language models in drug discovery and development: From disease
689 mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*, 2024.
- 690 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuur-
691 mans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex
692 reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.
- 693
- 694 Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang,
695 Ming Qin, Kehua Feng, Jike Wang, et al. Advancing biomolecular understanding and design
696 following human instructions. *Nature Machine Intelligence*, pp. 1–14, 2025.
- 697 Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. Kg4diagnosis: A hierarchical multi-agent llm
698 framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program*
699 *on AI for Medicine and Healthcare*, pp. 195–204. PMLR, 2025.

700
701

702 A APPENDIX

703 A.1 DETAILED GRPO OBJECTIVE AND IMPLEMENTATION

704 For a training input x , we sample G candidate responses $\{y_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{old}}$, and
 705 optimize the updated policy π_{θ} via the following clipped surrogate objective:

706
$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot|x)} \frac{1}{G} \sum_{i=1}^G \left[\min(p_i(\theta)A_i, \text{clip}(p_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i) \right. \\ \left. - \beta \mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{\theta_{ref}}) \right], \tag{12}$$

707 where $p_i = \frac{\pi_{\theta}(y_i | x)}{\pi_{\theta_{old}}(y_i | x)}$ is the probability ratio.

708 The normalized advantage A_i is computed within the sampled group:

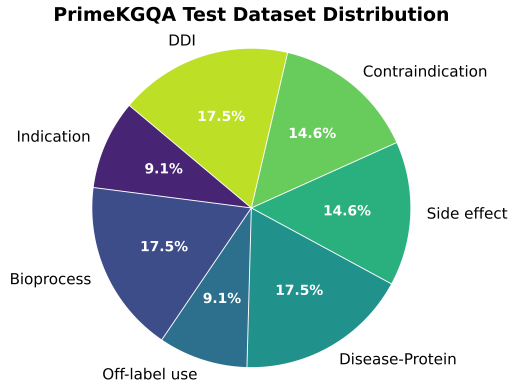
709
$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}, \tag{13}$$

710 with r_i the reward of the i -th response. The KL penalty term \mathbb{D}_{KL} stabilizes training against the
 711 reference policy $\pi_{\theta_{ref}}$, where ϵ is the clipping threshold and β the penalty weight.

712 Here, we provide a detailed breakdown of the components in the objective function. In this formula-
 713 tion, \mathcal{D} is the distribution of training prompts x , and G is the number of candidate responses sampled
 714 per prompt. The policy being optimized is π_{θ} , with parameters θ , while $\pi_{\theta_{old}}$ is a fixed, older version
 715 of the policy used for sampling. The reference policy, $\pi_{\theta_{ref}}$, is used to regularize π_{θ} . The term $p_i(\theta)$
 716 is the importance sampling ratio for the i -th response, and A_i is the normalized advantage of that
 717 response, calculated using the scalar reward r_i . The clip(\cdot) function constrains the probability ratio
 718 to stabilize training. The Kullback-Leibler (KL) divergence, $\mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{\theta_{ref}})$, acts as a penalty term,
 719 with its strength controlled by the coefficient β . Finally, ϵ is the clipping hyperparameter for the
 720 surrogate objective.

721 A.2 EVALUATION DATASETS

722 The distribution of the PrimeKGQA test dataset across various tasks is presented in Figure 4. The
 723 number of test samples for each task is proportional to the volume of corresponding data in the
 724 source KG, thus aligning the evaluation with the inherent data distribution of the KG. Table 3
 725 presents a quantitative overview of our test data, illustrating the statistical distribution of samples
 726 between the primary test set and the generalization benchmarks. The summary specifically delin-
 727 eates the volume and variety of type of questions within each data set, establishing a clear and
 728 comprehensive foundation for our performance evaluation.



730 Table 3: Comprehensive statistics for our evalua-
 731 tion data, detailing the distribution of question
 732 types and counts across the primary test set and
 733 generalization benchmarks.

Dataset	Question Type	Number
PrimeKGQA	MCQ	1710
BiomixQA	T/F Question	498
BioASQ	MCQ	246
MEDDDX	MCQ	245

734 Figure 4: The distribution of the PrimeKGQA
 735 test dataset across various tasks.

A.3 CASE STUDY AND RELATED PROMPTS

To provide a concrete illustration of our methodology, we present a detailed case study (Figure 5). The process begins with the initial problem, from which we extract a relevant reasoning path within the KG. This path serves as the foundation for generating an initial, verbose response using our CoT generation prompt (Figure 6). Subsequently, this preliminary answer undergoes a refinement phase, where a specialized CoT pruning prompt is applied to revise the essential information and produce the final, concise answer (Figure 7). Throughout these figures, we use color-coding to highlight the transformation. Specifically, the text marked in red indicates superfluous content that deviates from a true inferential path. Such content is revised because it is either superfluous or fails to demonstrate true reasoning. This failure is evident when the model abandons the step-by-step derivation, instead opting to provide external facts directly or flagging its input with terms like "Additional knowledge". In contrast, the content marked in green represents the final, polished output, which is composed of the valid inferential chain that has been preserved and, in some cases, augmented with further elaborations to enhance logical coherence and completeness.

A.4 DETAILED RESULTS

Our experimental results comprehensively validate the effectiveness and robustness of the proposed Bio-KCoT framework. The ablation study, detailed in Table 4, demonstrates that our method consistently and substantially outperforms both the base model and a strong distilled CoT baseline across all task categories and difficulty levels on the PrimeKGQA benchmark. This highlights the significant advantage of integrating structured knowledge into the reasoning process. Furthermore, as shown in Table 5, the performance gains are not confined to our primary dataset; the model maintains its superiority on several external biomedical benchmarks. This confirms that the enhanced reasoning capabilities generalize effectively to unseen data and diverse problem types, establishing Bio-KCoT as a powerful and broadly applicable framework for complex biomedical question answering.

Table 4: Ablation study on PrimeKGQA. Experiments are performed across all tasks with the 4B and 8B model variants, in comparison with the distilled CoT baseline.

Model	Basic			Medium				Hard			All Avg.
	Indication	Bio-process	Avg.	Off-label use	Disease-Protein	Side effect	Avg.	Contraindication	DDI	Avg.	
Qwen3-4B	0.768	0.523	0.606	0.710	0.573	0.616	0.618	0.438	0.503	0.473	0.568
Qwen3-4B (distilled CoT)	0.819	0.690	0.734	0.800	0.687	0.808	0.755	0.720	0.560	0.633	0.710
Bio-KCoT(4B)	<u>0.897</u>	<u>0.773</u>	<u>0.815</u>	<u>0.890</u>	0.743	0.912	<u>0.835</u>	<u>0.804</u>	<u>0.610</u>	0.698	<u>0.786</u>
Qwen3-8B	0.845	0.653	0.749	0.671	0.600	0.516	0.586	0.452	0.537	0.498	0.601
Qwen3-8B (distilled CoT)	0.869	0.753	0.793	0.851	<u>0.747</u>	0.828	0.798	0.768	0.597	0.674	0.773
Bio-KCoT(8B)	0.901	0.807	0.839	0.930	0.793	<u>0.908</u>	0.864	0.848	0.643	0.736	0.816

Table 5: Generalization study on three biomedical benchmarks (BiomixQA, BioASQ, and MEDDDX). Performance is evaluated using the 4B and 8B model variants.

Model	BiomixQA	BioASQ	MEDDDX
Qwen3-4b	0.472	0.659	0.426
Bio-KCoT(4b)	0.512(+4.0)	0.727(+6.8)	0.473(+4.7)
Qwen3-8b	0.515	0.759	0.507
Bio-KCoT(8b)	0.573(+5.8)	0.807(+4.8)	0.567(+6.0)

A.5 ERROR ANALYSIS

While the Bio-KCoT framework demonstrates substantial advancements in reasoning performance, a systematic error analysis is essential for diagnosing its inherent limitations and guiding future it-

810 erations. This section presents a detailed error analysis of failures arising during the generation and
811 execution of reasoning chains. Errors are traced from the final answers (highlighted in red) back
812 through the CoT generation and execution steps (Figure 8). This process helps to illuminate the
813 model’s current limitations, and informing pathways for future refinement. The analysis indicates
814 that most errors do not arise from flawed logical deduction but are instead attributable to the limited
815 knowledge capacity of smaller models, which may lead to errors when recalling specific facts or
816 specialized knowledge points. Consequently, even when the reasoning process is internally consist-
817 ent, the final outputs can contain inaccuracies or misconceptions, highlighting the importance of
818 combining reasoning analysis with knowledge verification in future iterations.

819 B ETHICS STATEMENT

820 This work complies with the ICLR Code of Ethics. Our study did not involve human participants or
821 animal testing. All datasets, including PrimeKGQA, were obtained and used in accordance with rel-
822 evant guidelines, ensuring respect for data usage policies and privacy considerations. We carefully
823 mitigated potential biases and avoided discriminatory effects during the research process. No per-
824 sonal or identifiable information was utilized, and no experiments were carried out that might raise
825 concerns regarding security or confidentiality. Throughout the study, we have upheld principles of
826 transparency, responsibility, and research integrity.

827 C REPRODUCIBILITY STATEMENT

828 To promote transparency and ensure reproducibility, we have provided all relevant code in the Sup-
829 plementary Material. The manuscript offers a comprehensive description of the experimental setup,
830 including the training pipeline, model architectures, hyperparameter configurations. These materials
831 are intended to enable independent verification of our findings and to support subsequent research
832 that builds upon our methodology.

833 D THE USE OF LARGE LANGUAGE MODELS

834 Large Language Models were employed as versatile assistive tools in this research. They were pri-
835 marily used to support and refine the preparation of this manuscript, including correcting grammar,
836 enhancing clarity, and restructuring sentences to improve overall readability. LLMs also assisted in
837 generating foundational scripts for data preprocessing tasks such as cleaning, formatting, and visu-
838 alization, which were subsequently reviewed and adjusted by the authors to ensure correctness and
839 consistency with research objectives.

840 In addition, LLM APIs were invoked during the CoT generation and pruning stages, where they
841 contributed to producing candidate reasoning traces and improving their quality through refinement.
842 Across all these applications, the outputs provided by LLMs were critically assessed and edited by
843 the authors. At no stage did the models substitute for human intellectual contributions; rather, they
844 served to accelerate and augment different phases of the research workflow.

845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

Question

Which disease can be treated with Lisinopril?

- A. qualitative or quantitative protein defects in neuromuscular diseases
- B. intellectual disability-hypotonia-brachycephaly-pyloric stenosis-cryptorchidism syndrome
- C. myocardial infarction
- D. craniofrontonasal dysplasia-Poland anomaly syndrome

Extracted path from KG

Lisinopril targets **ACE** and **REN**.

ACE is involved in the pathway: **Metabolism of Angiotensinogen to Angiotensins**, which is part of the broader **Peptide hormone metabolism pathway**.

ACE is associated with a wide range of diseases, including neurological disorders (e.g., schizophrenia, Alzheimer disease, cerebellar degeneration, anxiety disorder, bipolar disorder, depression), cardiovascular diseases (e.g., coronary artery disease, heart failure, **myocardial infarction**, stroke, atrial fibrillation), kidney diseases (e.g., chronic kidney disease, diabetic nephropathy, renal tubular dysgenesis), various cancers (e.g., lung cancer, liver cancer, gastric cancer, prostate cancer, salivary gland adenocarcinoma), metabolic and systemic disorders (e.g., substance abuse, hypotensive disorder, fatty liver disease), autoimmune and infectious conditions (e.g., celiac disease, staphylococcal infection, severe acute respiratory syndrome), and other conditions such as autism spectrum disorder and drug psychosis.

REN is involved in the same pathway: **Metabolism of Angiotensinogen to Angiotensins**, under **Peptide hormone metabolism**.

REN is associated with diseases including neurological and psychiatric disorders (e.g., anxiety disorder, depression, bipolar disorder, dysthymic disorder, neurotic disorder), cardiovascular and renal conditions (e.g., congestive heart failure, **malignant hypertension**, kidney disease, nephrosis, renal tubular dysgenesis, interstitial nephritis), metabolic and systemic disorders (e.g., anemia, osteoporosis, substance withdrawal syndrome, hypotensive disorder), autoimmune and skin diseases (e.g., psoriasis, pustulosis palmaris et plantaris, parapsoriasis), and genetic disorders affecting electrolyte balance (e.g., Bartter disease, Liddle syndrome, hyperaldosteronism, familial juvenile hyperuricemic nephropathy).

CoT Generation

Step1: The question asks which disease can be treated with Lisinopril, a drug known to target ACE (angiotensin-converting enzyme) and REN (renin). Both of these proteins play central roles in the Renin-Angiotensin-Aldosterone System (RAAS), which regulates blood pressure, vascular tone, and fluid balance. The key pathway involved is the metabolism of angiotensinogen to angiotensins, a subset of peptide hormone metabolism. Dysregulation of this pathway often leads to cardiovascular and renal pathologies, including hypertension, heart failure, and myocardial infarction.

Step2: **Knowing Lisinopril is an ACE inhibitor, I can now use the provided "Additional knowledge".** ACE is involved in the "Metabolism of Angiotensinogen to Angiotensins" pathway. This is the core of the Renin-Angiotensin-Aldosterone System (RAAS), a critical hormonal cascade that regulates blood pressure and fluid balance. By inhibiting ACE, Lisinopril prevents the conversion of angiotensin I to angiotensin II. Angiotensin II is a potent vasoconstrictor (it narrows blood vessels) and stimulates the release of aldosterone, which causes the body to retain salt and water. Therefore, blocking its production leads to vasodilation (widening of blood vessels) and a reduction in blood volume, both of which lower blood pressure and reduce the workload on the heart.

Step3: Now, I can evaluate each option in the context of Lisinopril's mechanism: A. qualitative or quantitative protein defects in neuromuscular diseases: These are typically genetic disorders affecting muscle function, like muscular dystrophies. The RAAS pathway is not the primary driver of this pathology. An ACE inhibitor would not correct underlying protein defects. B. intellectual disability-hypotonia-brachycephaly-pyloric stenosis-cryptorchidism syndrome: This is a complex, multi-system congenital syndrome. Its cause is genetic, not related to the hormonal regulation of blood pressure. Treatment would be supportive, and Lisinopril would not address the root causes. C. myocardial infarction (MI): This is a heart attack, a major cardiovascular event. **The "Additional knowledge" explicitly states that ACE is associated with... cardiovascular diseases (e.g., ... myocardial infarction). This is a direct link.** Clinically, ACE inhibitors like Lisinopril are a cornerstone of post-MI therapy because they reduce the strain on the damaged heart, lower blood pressure, and help prevent negative

Figure 5: A detailed case study illustrating our three-stage methodology. The example demonstrates the transformation of a response, starting from the foundational KG path, to a verbose answer via CoT generation, and concluding with a concise, refined answer after CoT pruning.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

structural changes (remodeling) in the heart muscle, thereby improving survival and reducing the risk of future heart failure. The drug's mechanism aligns perfectly with this indication. D. craniofrontonasal dysplasia-Poland anomaly syndrome: This is another congenital disorder involving developmental abnormalities of the head, face, and chest. It has no mechanistic connection to the RAAS pathway. **Step 4:** Based on the logical chain, Lisinopril's identity as an ACE inhibitor directly links it to the RAAS pathway. The function of this pathway is central to cardiovascular health. Of the choices provided, only myocardial infarction is a cardiovascular disease where the therapeutic effects of ACE inhibition are not only plausible but are a standard and evidence-based treatment. The provided knowledge directly confirms the association between the drug's target (ACE) and this specific disease.

CoT Pruning

Step 1: Alright, the question is about which disease Lisinopril can treat. Lisinopril—I recognize that name. It's an angiotensin-converting enzyme (ACE) inhibitor. ACE is a key enzyme in the renin-angiotensin-aldosterone system (RAAS). Lisinopril works by binding to the active site of ACE, which normally converts angiotensin I to angiotensin II. Angiotensin II is a potent vasoconstrictor and also stimulates aldosterone release, leading to sodium and water retention. By inhibiting ACE, Lisinopril reduces angiotensin II levels, causing vasodilation, lowering blood pressure, and decreasing fluid volume. But it also increases bradykinin because ACE breaks down bradykinin—so less degradation means more bradykinin, which promotes vasodilation through nitric oxide and prostaglandins. This makes me think Lisinopril is primarily used for cardiovascular conditions like hypertension or heart failure. Now, I need to match this mechanism to one of the diseases listed.

Step 2: Let me evaluate each option carefully, starting with A: "qualitative or quantitative protein defects in neuromuscular diseases." This sounds like disorders such as muscular dystrophy or myopathies, where there's a defect in proteins like dystrophin or components of the muscle membrane. These involve structural issues in muscle cells or neuromuscular junctions. Lisinopril affects systemic vascular tone and fluid balance through RAAS, but it doesn't target protein synthesis, repair genetic defects, or directly influence neuromuscular signaling. I recall that treatments for neuromuscular diseases focus on physical therapy, immunosuppressants, or gene therapies—not ACE inhibitors. So, this seems like a distractor; Lisinopril wouldn't address the core pathology here. I can rule out A confidently.

Step 3: Next, option B: "intellectual disability-hypotonia-brachycephaly-pyloric stenosis-cryptorchidism syndrome." That's a mouthful—it seems like a rare genetic syndrome, possibly linked to chromosomal deletions or mutations affecting neurodevelopment and physical features. Intellectual disability and hypotonia suggest brain involvement, while pyloric stenosis and cryptorchidism are structural anomalies. Lisinopril's mechanism is purely pharmacological, modulating blood pressure and cardiac load. It doesn't cross the blood-brain barrier effectively to affect cognition, nor can it correct congenital malformations like undescended testes or stomach narrowing. Genetic syndromes require targeted interventions, not broad-spectrum drugs like ACE inhibitors. I'm doubtful this could be correct; it feels unrelated to Lisinopril's known actions. So, B is likely not the answer.

Step 4: Option C: "myocardial infarction." This is a heart attack, where blood flow to part of the heart is blocked, causing tissue death. Post-MI, the RAAS gets hyperactivated as a stress response—angiotensin II drives inflammation, fibrosis, and ventricular remodeling, which worsens heart function over time. Lisinopril should counteract this by reducing angiotensin II, thus lowering afterload, improving coronary perfusion, and preventing pathological changes like wall thinning and dilation. It's a standard post-MI therapy to prevent heart failure. This seems plausible, but I should check the other option before concluding.

Step 5: Finally, option D: "craniofrontonasal dysplasia-Poland anomaly syndrome." This appears to be another genetic disorder, involving craniofacial and skeletal defects like fused sutures or missing chest muscles. Mutations in genes like EFNB1 disrupt cell migration during development. Lisinopril can't alter genetic coding or fix structural birth defects—it might even be risky in syndromes with potential renal or vascular abnormalities. Treatments involve surgery or specialized care, not RAAS modulation. I have no evidence linking ACE inhibitors to such conditions; this option seems like a distractor similar to B. So, D is incorrect.

Step 6: Now, confirming my hypothesis for C: Lisinopril's role in myocardial infarction isn't just theoretical—it's well-established. By inhibiting RAAS, it reduces infarct expansion, stabilizes the extracellular matrix, and lowers oxygen demand. The bradykinin buildup might also have cardioprotective effects. **Conclusion:** After eliminating distractors—A (neuromuscular, no RAAS link), B (genetic neurodevelopmental, not pharmacologically addressable), and D (genetic skeletal, requires surgical intervention)—only C, myocardial infarction, aligns with Lisinopril's mechanism and real-world clinical applications. Therefore, C is the correct answer.

Figure 5: (continued)

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

The prompt for Generation

You are given a single-choice biological reasoning question. Your task is to generate a detailed step-by-step reasoning process that can guide a model to select the correct answer.

Question: *{question}*
Options: *{options}*
Additional knowledge: *{p}*
True answer: *{true_answer}*

Now, write a natural, plausible reasoning process that could lead a knowledgeable person to choose the correct answers. Avoid simply referencing pre-established facts, and instead build your reasoning as a sequence of logical thoughts that connect concepts.
Begin by examining *{head_node_type}* biological basis, including key pathways, affected tissues, and pathological mechanisms. Then reason about which proteins among the options are likely to be functionally or mechanistically involved, based on their roles and interactions. Evaluate each option independently and identify all proteins that are plausible based on this reasoning.

Do not just state facts from the "Additional knowledge", integrate them into a coherent narrative, making the reasoning flow as though it arises naturally from your own deep biological understanding.

Finally, return your answer as a JSON dictionary with the reasoning in the field "cot_reasoning".
Output format: `{{"cot_reasoning": ["xxx"]}}`

Figure 6: The prompt structure used for CoT generation. This prompt guides the model to produce a detailed, step-by-step reasoning process that leads to an initial answer based on the extracted KG path.

The prompt for Pruning

Modify the scientific reasoning to meet the following requirements.
Please follow these instructions:

1. Keep the technical accuracy and depth of the original reasoning.
2. Delete the content that is not related to the question.
3. Add realistic thought processes: moments of doubt and elimination of distractors.
4. Ensure every answer option is explicitly considered and evaluated.

Here is an example: *{example question}*
Original output: *{example original_output}*
The model should rewrite like this: *{example rewrite_output}*

Now, try to rewrite the thinking process of the next question in the same way.
Question: *{question}*
Original output: *{original_output}*

[Output Format]
`<reasoning>\n ... \n</reasoning>`

Figure 7: The prompt structure used for CoT pruning. This prompt instructs the model to refine the initial, verbose reasoning by eliminating redundancies, correcting logical failures, and distilling the core inferential steps into a final, concise answer.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

Error cases

Question

Which disease is contraindication for Carbenicillin?

- A. non-suppurative otitis media
- B. optic nerve glioma
- C. perinatal necrotizing enterocolitis
- D. pregnancy associated osteoporosis

Error answer

Step 1: Okay, the question is about contraindications for Carbenicillin. First, I need to recall what Carbenicillin is. It's a beta-lactam antibiotic, specifically a carboxypenicillin, used primarily against Gram-negative bacteria like *Pseudomonas aeruginosa*. Contraindications are conditions where the drug should not be used due to potential harm. I remember Carbenicillin has known risks, especially with prolonged use or in vulnerable populations. The big ones are nephrotoxicity (kidney damage) and neurotoxicity (like seizures or nerve issues). Also, since it's a penicillin derivative, there's a risk of allergic reactions, but that's more of a general caution than a specific contraindication.

Step 2: Now, I need to evaluate each disease option systematically. Let's take them one by one, thinking about mechanisms and potential conflicts.
 A. Non-suppurative otitis media: This is a type of middle ear infection without pus, often viral. Antibiotics like Carbenicillin might be used if bacterial, but is it a contraindication? I'm doubtful. Non-suppurative cases are usually viral, so antibiotics aren't first-line. Even if bacterial, Carbenicillin's risks (nephrotoxicity, neurotoxicity) aren't directly tied to ear infections. The main risks here are misuse leading to resistance or side effects, not a specific prohibition. This seems like a distractor—it's not a core contraindication, so I can probably rule it out.
 B. Optic nerve glioma: This is a tumor in the optic nerve, often seen in children. Hmm, why might Carbenicillin be contraindicated here? I recall that beta-lactam antibiotics like Carbenicillin can cause neurotoxicity, including optic neuritis or nerve damage. If someone already has an optic nerve glioma, giving this drug could worsen the condition by exacerbating inflammation or toxicity in that nerve.

That makes sense—using a neurotoxic drug in a patient with pre-existing optic nerve damage could lead to vision loss or accelerated tumor progression. This feels plausible, but I should check the other options before deciding.
 C. Perinatal necrotizing enterocolitis: This is a serious intestinal condition in newborns. Carbenicillin is sometimes used in neonates for infections, but is it contraindicated? I'm skeptical. While antibiotics can contribute to NEC in some cases (e.g., by disrupting gut flora), Carbenicillin isn't specifically contraindicated. Its risks like nephrotoxicity might be more relevant in older patients than in neonates. Also, NEC management focuses on supportive care and avoiding certain drugs, but Carbenicillin isn't a typical no-go. This seems like another distractor—no strong link to the drug's specific dangers.
 D. Pregnancy associated osteoporosis: This is bone loss in pregnant women, often due to hormonal changes. Carbenicillin is a penicillin, and penicillins can be used safely in pregnancy (like for bacterial infections), but they're not known to cause osteoporosis or worsen it. The drug's metabolism doesn't involve bone pathways, and there's no evidence it affects calcium or vitamin D. While all drugs should be used cautiously in pregnancy, this isn't a specific contraindication. I can eliminate this one—it's unrelated.

Step 3: Going back to option B, I need to test the hypothesis. If optic nerve glioma is a contraindication, why? Carbenicillin's neurotoxicity can manifest as optic neuritis, which might mimic or worsen a glioma. In clinical practice, avoiding neurotoxic drugs in patients with pre-existing optic nerve pathology is prudent to prevent irreversible damage. For instance, in conditions like multiple sclerosis, neurotoxic drugs are avoided for similar reasons. This aligns with Carbenicillin's profile, making B the strongest candidate.

Step 4: Final confirmation. After eliminating A, C, and D, only B has a direct mechanistic link: Carbenicillin's potential to cause neurotoxicity could harm the optic nerve in a glioma patient. No other option has this level of risk association. So, the contraindication stands.

Figure 8: An illustrative example of the error analysis process for a failure case. The error is traced from the final answer back through the reasoning chain, revealing that the failure originates from a knowledge deficit rather than a flaw in logical deduction.