

# RARE: Retrieval-Augmented Reasoning Enhancement for Large Language Models

Anonymous ACL submission

## Abstract

This work introduces **RARE (Retrieval-Augmented Reasoning Enhancement)**, a versatile extension to the mutual reasoning framework (rStar), aimed at enhancing reasoning accuracy and factual integrity across large language models (LLMs) for complex, knowledge-intensive tasks such as medical and common-sense reasoning. RARE incorporates two innovative actions within the Monte Carlo Tree Search framework: **(A6)**, which generates search queries based on the initial problem statement, performs information retrieval using those queries, and augments reasoning with the retrieved data to formulate the final answer; and **(A7)**, which leverages information retrieval specifically for generated sub-questions and re-answers these sub-questions with the relevant contextual information. Additionally, a Retrieval-Augmented Factuality Scorer is proposed to replace the original discriminator, prioritizing reasoning paths that meet high standards of factuality. Experimental results with LLaMA 3.1 show that RARE enables open-source LLMs to achieve competitive performance with top closed-source models like GPT-4 and GPT-4o. This research establishes RARE as a scalable solution for improving LLMs in domains where logical coherence and factual integrity are critical <sup>1</sup>.

## 1 Introduction

Question answering (QA) is a cornerstone task in natural language processing that involves generating answers to questions posed in natural language. QA spans a broad spectrum of domains and types, ranging from open-domain QA (Yang et al., 2018; Kwiatkowski et al., 2019) to more specialized areas like medical QA (Jin et al., 2021; Cao et al., 2011). The overwhelming volume and complexity of medical information necessitate medical QA, which benefits many downstream tasks such as medical

education, clinical decision support, and patient care optimization (Cai et al., 2023; Liu et al., 2023; Jin et al., 2024b).

Medical QA represents a unique and demanding subset of QA, requiring models to navigate intricate medical knowledge, interpret clinical scenarios, and select correct and contextually appropriate options (Singhal et al., 2023b; Wu et al., 2024). Similar to general domain QA, Medical QA requires structured multi-step reasoning, where answers emerge from various sequential steps. Take Figure 1 as an example, to find appropriate treatment given patient information, the QA model should first identify patient conditions (colored in red, e.g., chief complaint and past conditions), then analyze contributing factors and diagnose the disease (colored in blue), and determine appropriate evidence-based interventions in the final step (colored in yellow). Without such structured multi-step reasoning, it would be challenging to arrive at an accurate and contextually relevant answer for such a complex medical question.

Moreover, Medical QA presents two non-trivial challenges that distinguish it from general-domain QA. First, Medical QA depends heavily on domain-specific knowledge that is not always available within pre-trained models, necessitating knowledge-based retrieval from external sources (Xiong et al., 2024a). Figure 1 is an example which involves specific medical terms such as *allergic conjunctivitis*. In addition, medical knowledge evolves rapidly, and new treatments or updated guidelines may not be included in the model’s pre-trained corpus. For example, newer drugs (like epinastine hydrochloride for allergic conjunctivitis) may be recommended by recent guidelines but absent in older pre-trained models. Second, Medical QA encompasses a wide variety of question types, including not only multi-step reasoning and knowledge-based retrieval as previously mentioned, but also composite questions requiring iterative evi-

<sup>1</sup>The anonymous version of our code can be accessed at: <https://anonymous.4open.science/r/RARE-EBEB>

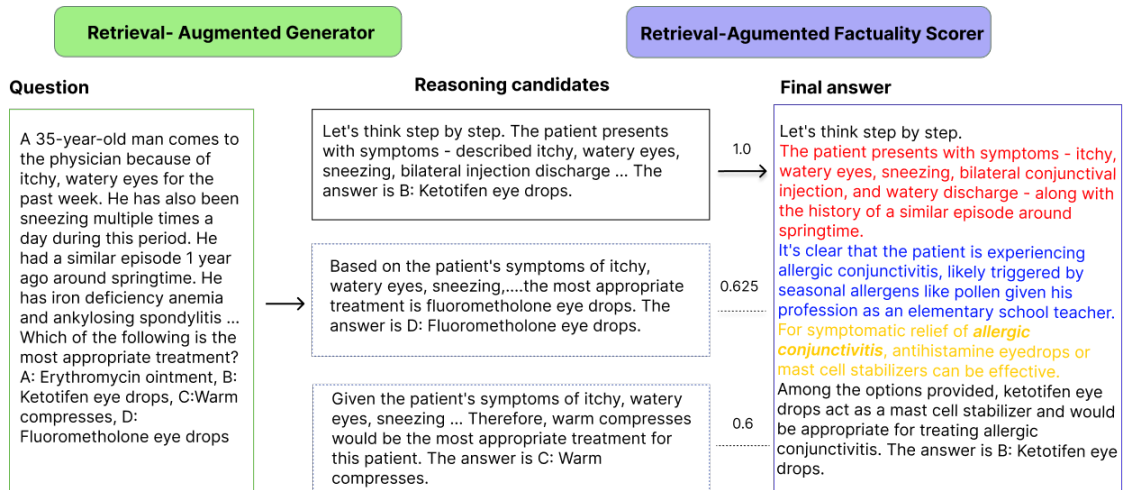


Figure 1: Overview of our reasoning process, which combines generation and factuality scoring. (1) A retrieval-augmented generator produces multiple candidate reasoning paths using Monte Carlo Tree Search (MCTS); (2) a retrieval-augmented factuality scorer evaluates the factual accuracy of each reasoning path; (3) the trajectory with the highest factuality score is selected as the final answer.

dence retrieval, where they demand retrieval of relevant knowledge at each reasoning step to ensure accuracy and relevance throughout the process.

In parallel, Commonsense Question Answering shares similar complexities with Medical QA, particularly in its reliance on structured multi-step reasoning and iterative evidence retrieval. While Medical QA draws heavily on domain-specific knowledge, Commonsense QA focuses on leveraging a model’s understanding of general world knowledge and logical connections to answer questions that are often indirect or abstract. For example, tasks like StrategyQA (Geva et al., 2021) require models to infer hidden relationships and execute multi-hop reasoning, akin to diagnosing a condition in Medical QA (Trivedi et al., 2023; Bauer et al., 2018; Chen et al., 2020). This similarity in reasoning processes across both domains underscores the importance of designing frameworks that can adapt to and optimize multi-step reasoning workflows, irrespective of the domain.

In this paper, we propose Retrieval-Augmented Reasoning Enhancement (RARE) to handle aforementioned challenges. We built upon rStar (Qi et al., 2024) where a language model generates reasoning steps and another verifies them, improving accuracy without fine-tuning or superior models. To answer knowledge-based questions, RARE designed a new action A6, which generates multiple search queries based on the question and retrieves relevant documents. To answer composite questions, we add action A7, which refines sub-questions, retrieves targeted information, and updates the next step. RARE applies the Monte Carlo Tree Search (MCTS) algorithm to select the best ac-

tion path that leads to the final answer. In addition, RARE is complemented by Retrieval-Augmented Factuality Scorer (RAFS) that evaluates and ranks reasoning paths for factual accuracy.

We applied RARE and other baselines on 3 medical QA tasks and 3 general domain QA tasks. Results show that RARE significantly enhances accuracy across various LLMs, enabling the open-source LLMs (LLaMA 3) to achieve competitive performance with top closed-source LLMs like GPT-4o. Our contributions are as follows:

- Formulating Medical QA as Multi-Step Reasoning:** We build upon the rStar framework to model medical QA as a structured multi-step reasoning task, addressing the complexity and sequential nature of medical queries.
- Novel Retrieval Actions:** We introduce two retrieval-augmented actions within the MCTS framework, enabling the integration of real-time, context-specific information to enhance reasoning accuracy and relevance.
- Retrieval-Augmented Factuality Scorer:** We propose a Retrieval-Augmented Factuality Scorer to evaluate and rank reasoning paths, ensuring they maintain both logical coherence and factual reliability throughout the reasoning process.

## 2 Preliminaries

This section introduces the foundational concepts and notations used in this work, focusing on the Monte Carlo Tree Search (MCTS) algorithm and the rStar framework (Qi et al., 2024), which serve as the basis for our proposed RARE method.

## Actions

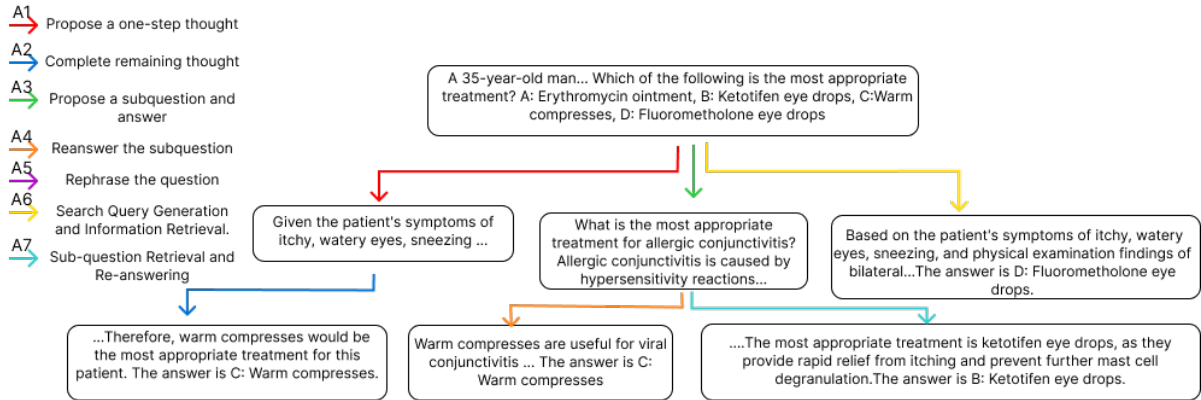


Figure 2: An example to illustrate the process of retrieval-augmented generator. Highlighted nodes from top to bottom constitute a complete reasoning trace. Given a question, MCTS augments the LLM to explore a rich, human-like reasoning action space and generate the next steps based on the current state.

## 2.1 Monte Carlo Tree Search (MCTS)

Monte Carlo Tree Search (MCTS) is a decision-making algorithm that incrementally builds a search tree by simulating outcomes to estimate action values, making it effective for complex reasoning tasks (Browne et al., 2012). It operates through iterative selection, expansion, simulation, and backpropagation phases, balancing exploration and exploitation via the Upper Confidence Bound applied on Trees (UCT). MCTS enables adaptive strategy refinement in large search spaces where direct computation is infeasible. A detailed explanation of MCTS and its implementation can be found in the Appendix A5.

## 2.2 Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers

Building upon MCTS, (Qi et al., 2024) proposed rStar, a framework augments MCTS with a diverse set of reasoning actions. This enhancement is designed to improve exploration of the solution space in complex reasoning tasks by allowing more dynamic and human-like reasoning pathways. Traditional approaches, such as Chain of Thought (CoT) reasoning (Wei et al., 2022) or self-consistency (Wang et al., 2022), often rely on a single action type, which can limit the diversity and effectiveness of generated solutions. In contrast, rStar incorporates five distinct actions that enable more adaptive exploration:

**A1: Propose a One-Step Thought.** This action generates the next reasoning step based on previous steps, allowing the LLM to build the solution incrementally.

**A2: Propose Remaining Thought Steps.** This action enables the LLM to produce all remaining reasoning steps in one inference, similar to CoT, for simpler questions.

**A3: Generate Next Sub-question and Answer.** This action decomposes the main problem into a sequence of sub-questions, each solved in turn.

**A4: Re-answer Sub-question.** This action allows the LLM to re-answer a previously generated sub-question, increasing accuracy by using few-shot prompting.

**A5: Rephrase Question/Sub-question.** This action rephrases the question to clarify conditions and reduce misunderstandings, enhancing the LLM’s interpretation of the problem.

## 3 Methodology

### 3.1 Overview of RARE Framework

Inspired by the generator-discriminator structure of rStar (Qi et al., 2024), RARE introduces a retrieval-augmented generator and a factuality scorer to enhance reasoning accuracy and factual integrity in large language models. As illustrated in Figure 1, RARE operates in two main stages:

**Candidate Generation with Retrieval-Augmented Generator:** The retrieval-augmented generator builds on the MCTS-based rStar self-generator, incorporating two new retrieval-augmented actions that dynamically fetch relevant external information. These actions improve the relevance and accuracy of candidate reasoning paths by integrating contextually enriched knowledge into intermediate reasoning steps, especially for complex questions, illustrated in Figure 2.

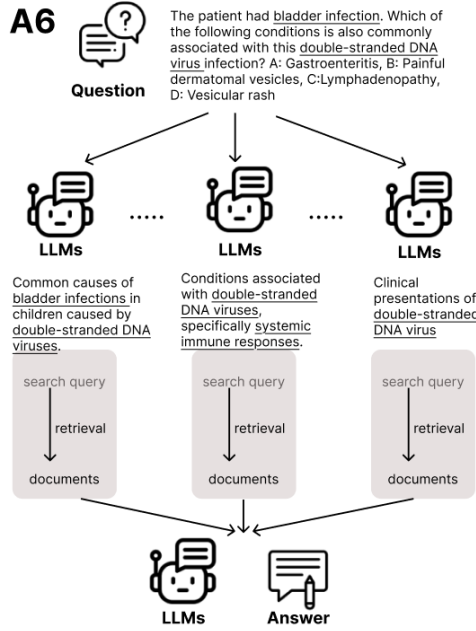


Figure 3: Illustration of the proposed retrieval-augmented action A6 in RARE: Given a question, LLMs generate search queries and retrieve relevant documents to construct a contextually enriched final answer. Key entities are underlined.

**Factuality Evaluation with Retrieval-Augmented Factuality Scorer:** Replacing the discriminator in rStar, the Retrieval-Augmented Factuality Scorer evaluates each candidate trajectory’s factual reliability. This scorer verifies the alignment of intermediate reasoning steps with retrieved evidence, assigning a factuality score that reflects the trajectory’s consistency with external knowledge. The trajectory with the highest factuality score is selected as the final answer, prioritizing the most factually supported reasoning path. This selection ensures coherence and factual alignment, enhancing response reliability.

Through these stages, RARE systematically integrates retrieval-based evidence into the reasoning process, optimizing both reasoning coherence and factual accuracy. This approach makes RARE well-suited for knowledge-intensive tasks, such as commonsense and medical reasoning.

### 3.2 Retrieval-Augmented Generator

Traditional retrieval-augmented generation methods often rely on a single retrieval step before generating responses, which limits their effectiveness in complex reasoning tasks. However, recent advances in medical RAG have demonstrated the importance of **iterative retrieval** in improving reasoning quality. i-MedRAG (Xiong et al., 2024b) highlights that multi-turn retrieval, where an LLM

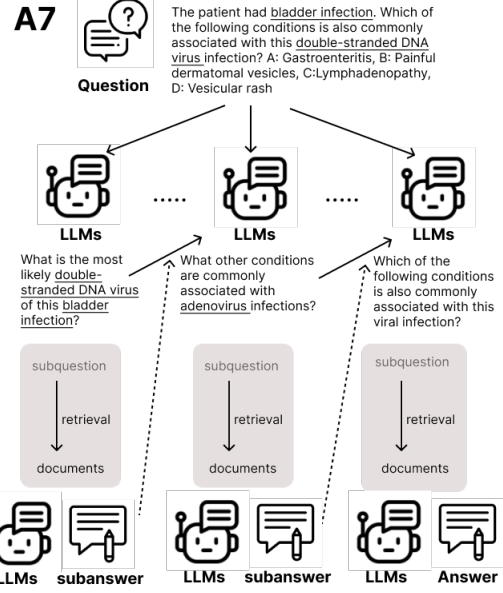


Figure 4: Illustration of the proposed retrieval-augmented action A7 in RARE: LLMs decompose the question into sub-questions, perform retrieval for each sub-question, and re-answer them based on the retrieved documents. The final sub-question is a rephrased version of the original question, so the sub-answer to this final sub-question also serves as the answer to the original question. In comparison with previous figure, we can find that A6 tends to use existing entity from the main question where A7 tends to use additional entity from previous subanswer.

generates follow-up queries dynamically, significantly enhances response accuracy by refining its knowledge base. Inspired by this approach, we introduce two new retrieval-augmented actions into the original rStar self-generator (Qi et al., 2024), transforming it into a **Retrieval-Augmented Generator**, as illustrated in Figures 3 and 4. These retrieval-augmented actions enable the generator to dynamically incorporate external knowledge during intermediate reasoning steps, improving both the contextual relevance and factual accuracy of generated responses.

**A6: Search Query Generation and Information Retrieval.** In this step, the LLM generates targeted search queries from the original question and retrieves relevant information. This additional context is then used to enrich the reasoning path, enabling the model to deliver a more comprehensive and contextually grounded final answer.

**A7: Sub-question Retrieval and Re-answering.** Unlike Action A6, which centers on the main question, this action targets sub-questions generated by Action A3. For each sub-question, the model fetches relevant information and re-answers accordingly. By refining these intermediary steps,



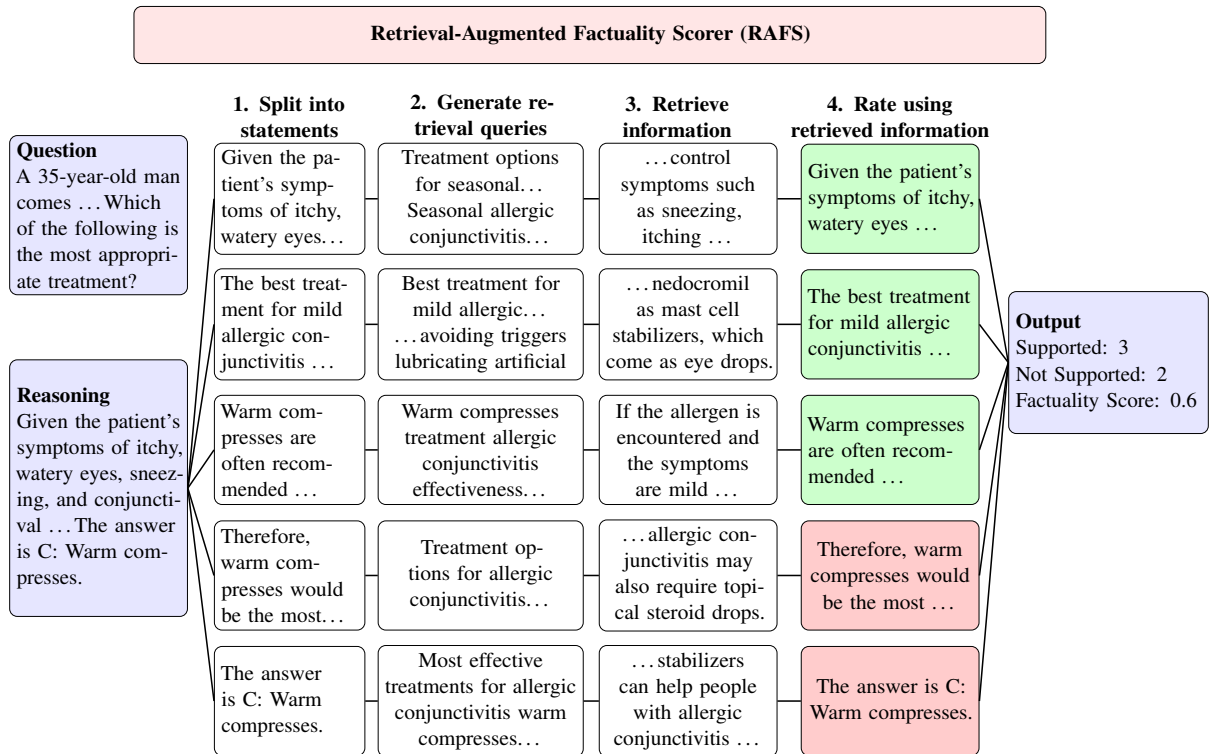


Figure 5: The RAFS assesses the factual accuracy of reasoning paths in four steps. (1) Split into sentences: The reasoning is divided into individual statements. (2) Generate retrieval queries: For each statement, an LLM generates multiple queries aimed at retrieving relevant information. (3) Retrieve information: The retrieval system gathers supporting information based on these queries. (4) Rate using retrieved information: Each statement is evaluated against the retrieved information and labeled as **Supported** or **Not Supported**. The final output includes a factuality score, calculated as the proportion of supported statements, which aids in selecting the most factually reliable reasoning path.

A7 improves the reasoning chain’s coherence and factual reliability, resulting in more accurate overall outcomes.

By incorporating these retrieval-augmented actions, the generator can explore a wider range of possible solutions, leading to reasoning paths that are both logically coherent and enriched with external knowledge. This upgrade transforms the generator into a retrieval-augmented generator, enabling RARE to better handle complex, knowledge-intensive reasoning tasks. We draw on a diverse corpus—PubMed, StatPearls, medical textbooks, and Wikipedia—for authoritative, up-to-date information. For efficient retrieval, we employ the ColBERT model (Khattab and Zaharia, 2020), a dense retrieval approach optimized for passage-level retrieval, enabling fine-grained token-level matching to enhance the precision of retrieved information. Additional details on the retrieval corpus and model can be found in Appendix A4.

### 3.3 Retrieval-Augmented Factuality Scorer (RAFS)

Inspired by the Search Augmented Factuality Evaluator (SAFE) (Wei et al., 2024), which combines an LLM (GPT-3.5-turbo) with Google Search to evaluate the factuality of responses, RARE introduces the **Retrieval-Augmented Factuality Scorer**. RAFS adapts this approach by replacing GPT-3.5-turbo with LLaMA 3.1 and Google Search with a corpus index retrieval system containing both general-domain knowledge (Wikipedia) and medical-domain resources (PubMed, StatPearls, and Medical Textbooks). This adaptation enhances domain specificity and accessibility for tasks requiring specialized knowledge. To assess the factual accuracy of generated reasoning paths, RAFS evaluates each candidate trajectory in four systematic steps, as illustrated in Figure 5.

**Split into Statements:** Each reasoning path is divided into individual statements.

**Generate Retrieval Queries:** For each statement, RAFS employs an LLM to generate multiple retrieval queries designed to retrieve contextually

Model	Method	MedQA	MedMCQA	MMLU-M	Avg
LLaMA3.2 3B	CoT	52.63	49.82	57.67	53.37
	MedRAG	52.08	51.78	65.58	56.48
	i-MedRAG	60.88	53.60	66.76	60.41
	SC	56.09	50.85	58.49	55.14
	rStar	61.27	54.26	67.22	60.92
	RARE	<b>63.86</b>	<b>56.61</b>	<b>70.98</b>	<b>63.82</b>
LLaMA3.1 8B	CoT	61.51	55.15	71.63	62.76
	MedRAG	63.00	56.87	74.56	64.81
	i-MedRAG	73.61	61.61	78.42	71.21
	SC	64.73	56.35	72.73	64.60
	rStar	70.40	62.13	79.16	70.56
	RARE	<b>75.57</b>	<b>64.32</b>	<b>81.63</b>	<b>73.84</b>
LLaMA3.1 70B	CoT	76.67	68.75	81.72	75.71
	MedRAG	77.61	71.19	84.76	77.85
	i-MedRAG	82.40	72.38	86.69	80.49
	SC	79.49	70.19	82.73	77.47
	rStar	84.99	72.72	87.19	81.63
	RARE	<b>87.43</b>	<b>75.18</b>	<b>90.91</b>	<b>84.51</b>
Meditron 70B	CoT	51.69	46.74	64.92	54.45
Mixtral (8x7B)	CoT	64.10	56.28	74.01	64.80
GPT-3.5	CoT	65.04	55.25	72.91	64.40
GPT-4	CoT	83.97	69.88	89.44	81.10
GPT-4o Mini	CoT	73.29	66.17	84.30	74.59
GPT-4o	CoT	85.55	74.70	90.45	83.57

Table 1: Performance of RARE and baseline methods on three medical reasoning benchmarks: MedQA, MedMCQA, and MMLU-Medical. SC is self-consistency.

relevant evidence.

**Retrieve Information:** The retrieval system gathers documents or information that corresponds to each generated query.

**Rate Using Retrieved Information:** Each statement is compared against the retrieved evidence and labeled as either **Supported** or **Not Supported**, based on alignment with the information.

As shown in Figure 5, RAfs outputs a factuality score along with **Supported** or **Not Supported** labels for each statement. This scoring aids in selecting the most reliable reasoning path from multiple candidates, allowing RARE to prioritize responses that align closely with verified external knowledge.

## 4 Results

In this section, we evaluate the performance of our proposed method, RARE, on both medical reasoning and commonsense reasoning tasks using three large language models: LLaMA 3.2 3B Instruct, LLaMA 3.1 8B Instruct and LLaMA 3.1 70B Instruct (Dubey et al., 2024). Throughout our work, we may drop “Instruct”, but we are always referring to the “Instruct” versions. Detail settings of the experiments, descriptions of the evaluation tasks and baselines can be found in Appendix.

### 4.1 Performance on Medical Reasoning tasks

Table 1 shows the performance of RARE and various baseline methods on three challenging medical reasoning benchmarks: MedQA, MedM-

CQA, and MMLU-Medical. These datasets require not only complex reasoning but also a high degree of factual accuracy, making them suitable for evaluating the effectiveness of RARE’s retrieval-augmented reasoning approach. Across all model sizes—LLaMA3.2 3B, LLaMA3.1 8B, and LLaMA3.1 70B—RARE consistently outperforms baseline methods. On LLaMA3.1 8B, RARE achieves substantial gains, outperforming rStar by 5.17% on MedQA, 2.19% on MedMCQA, and 2.47% on MMLU-Medical. The performance improvement becomes more pronounced as model size increases, with RARE-enhanced LLaMA3.1 70B outperforming GPT-4o on MedQA (87.43% vs. 85.55%) and MMLU-Medical (90.91% vs. 90.45%), highlighting its competitive edge.

### 4.2 Performance on Commonsense Reasoning

Table 2 presents the performance of RARE compared to other methods and larger language models on commonsense reasoning benchmarks, including StrategyQA, CommonsenseQA, and Physical IQA. These datasets test a range of commonsense reasoning skills, with StrategyQA requiring more multi-step reasoning. RARE consistently outperforms baseline methods, including CoT, MedRAG, iMedRAG, SC and rStar, across both LLaMA3.1 8B and LLaMA3.1 70B models. For LLaMA3.1 8B, RARE achieves substantial improvements over rStar, with gains of 6.45% on StrategyQA, 4.26% on CommonsenseQA, and 2.87% on PIQA. On LLaMA3.1 70B, RARE further closes the gap with state-of-the-art proprietary models, achieving 85.74% on StrategyQA, 86.98% on CommonsenseQA, and 92.66% on PIQA, surpassing GPT-4o.

RARE’s retrieval-augmented reasoning method provides substantially larger gains over CoT on multi-step inference tasks (like StrategyQA) compared to commonsense-heavy tasks (such as CommonsenseQA), indicating its particular effectiveness in handling implicit, multi-hop reasoning. Specifically, RARE’s largest improvement over CoT is observed on StrategyQA(10.19%), where multi-step reasoning is crucial, suggesting that its retrieval-augmented reasoning enhancement method is particularly effective in handling implicit, multi-hop inference. In contrast, while RARE outperforms CoT on CommonsenseQA(7.22%), the gains are relatively smaller, indicating that tasks relying more on commonsense knowledge rather

Model	Method	SQA	CQA	PIQA	Avg
LLaMA3.1 8B	CoT	67.83	73.62	76.17	72.54
	MedRAG	66.08	74.45	78.67	73.07
	i-MedRAG	68.12	75.18	80.52	74.61
	SC	68.41	74.90	77.42	73.58
	rStar	71.57	76.58	83.04	77.06
	RARE	<b>78.02</b>	<b>80.84</b>	<b>85.91</b>	<b>81.59</b>
LLaMA3.1 70B	CoT	76.71	78.62	81.66	79.00
	MedRAG	75.54	82.23	86.07	81.28
	i-MedRAG	77.29	83.13	87.76	82.73
	SC	77.29	78.87	82.67	79.61
	rStar	81.80	83.66	89.27	84.91
	RARE	<b>85.74</b>	<b>86.98</b>	<b>92.66</b>	<b>88.46</b>
Claude-3 Haiku	CoT	69.58	67.40	82.32	73.10
Claude-3.5 Sonnet	CoT	76.86	74.12	89.39	80.12
GPT-4o Mini	CoT	78.60	82.31	88.41	83.11
GPT-4o	CoT	80.64	86.50	91.13	86.09

Table 2: Performance comparison on common sense reasoning tasks with various LLMs and reasoning methods, evaluated on StrategyQA (SQA), Common-senseQA (CQA), and Physical IQA (PIQA). SC is self-consistency.

than explicit step-by-step reasoning do not benefit as significantly from retrieval-augmented reasoning.

### 4.3 Ablation Study

Configuration	Accuracy
rStar	70.0
rStar + RAFS	71.4
rStar + A6	72.4
rStar + A7	71.2
rStar + A6 + A7	73.2
RARE (rStar + A6 + A7 + RAFS)	<b>74.8</b>

Table 3: Ablation study on RARE components, evaluated on 250 MedQA samples using LLaMA 3.1 8B.

We conduct an ablation study on 250 MedQA samples using the LLaMA 3.1 8B model to assess the contributions of each RARE component (Table 3). Starting with the baseline (rStar) at 70.0% accuracy, adding the RAFS improves reliability, increasing accuracy to 71.4%. Incorporating A6 and A7 further boosts accuracy to 72.4% and 71.2%, respectively. Combining both actions yields 73.2%, demonstrating their synergy in strengthening reasoning. The full RARE configuration, integrating rStar, A6, A7, and RAFS, achieves the highest accuracy (74.8%), highlighting the collective impact of retrieval and factuality scoring in enhancing reasoning accuracy.

### 4.4 Common Reasoning Paths Patterns

Exploring common paths leading to correct answers helps us understand and refine the reasoning patterns that most reliably produce accurate solu-

tions. Figures 8 and 9 illustrate the top 10 most common reasoning paths that lead to correct answers on MedQA and StrategyQA, respectively.

In MedQA, prominent paths such as A1 → A2, A3 → A2, and A1 → A6 constitute a large share of successful reasoning. More complex sequences like A3 → A7 → A3 also appear, demonstrating the generator’s capacity to explore multiple approaches adaptively.

A similar pattern occurs on StrategyQA, where paths like A1 → A2, A3 → A2, and A1 → A6 remain dominant. Still, simpler actions such as A6 and A2 also play a noticeable role, reflecting the distinct reasoning demands of this task.

Overall, both figures highlight the RARE’s flexibility in navigating a variety of reasoning strategies. They further illustrate the importance of both straightforward and more intricate paths in effectively tackling different categories of tasks.

### 4.5 Human Evaluation of the Retrieval-Augmented Factuality Scorer

Metric	Score (%)
Inter-Annotator Agreement	86.49
RAFS Alignment with Annotator 1	87.84
RAFS Alignment with Annotator 2	82.43
Average RAFS-Annotator Alignment	<b>85.14</b>

Table 4: Human evaluation of RAFS.

To assess the effectiveness of RAFS, we conducted a human evaluation comparing its factual assessments with expert judgments. We selected 10 medical questions from MedQA, generate corresponding responses by RARE with LLaMA 3 8B, segmenting their responses into around 100 statements. RAFS classified each statement as either **Supported** or **Not Supported**. Two medical experts independently labeled statements as **Correct** or **Incorrect**, aligning with RAFS’s **Supported** and **Not Supported** labels, respectively. Table 4 show high inter-annotator agreement 86.49% confirms consistency among experts. RAFS achieves strong alignment with human judgment 85.14% on average, demonstrating its reliability in factual assessment. The slightly lower agreement with Annotator 2 (82.43%) suggests some subjectivity in evaluating factual correctness. RAFS effectively prioritizes factually accurate reasoning paths, aligning closely with expert validation.

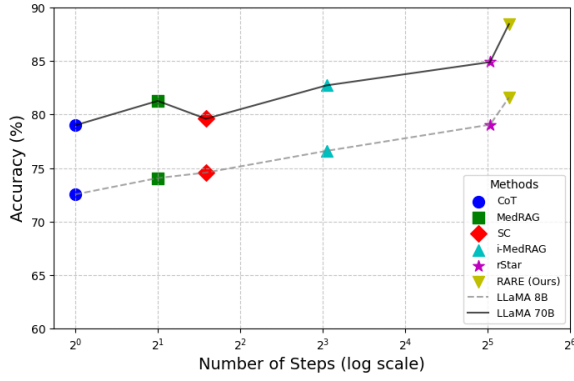


Figure 6: Accuracy vs. Number of Steps for Different Methods

## 4.6 Computational Analysis

Figure 6 illustrates the trade-off between inference cost and reasoning performance on average of common sense reasoning tasks. Each reasoning step corresponds to a single LLM call, with CoT requiring just one step, rStar averaging 32 steps, and RARE taking 38 steps due to additional retrieval-augmented steps. On average, rStar improves accuracy over CoT by 4.52%, while RARE achieves a 9.05% improvement, the highest among all methods. Although RARE requires roughly 6 more steps than rStar, its superior performance demonstrates that additional computation enhances reasoning robustness. This aligns with recent trends, such as OpenAI O1 and DeepSeek R1 (El-Kishky et al., 2025; Guo et al., 2025), which prioritize deeper reasoning over minimal latency, reinforcing that improved accuracy justifies the computational trade-off in complex reasoning tasks.

## 5 Related Work

**Prompting LLMs to reason** has been extensively explored through methods like Chain-of-Thought (Wei et al., 2022), aiming to improve inference via structured reasoning prompts. Advances include planning (Hao et al., 2023; Ding et al., 2023), problem decomposition (Zhou et al., 2022; Khot et al., 2022), and programming-based reasoning (Chen et al., 2022; Zhou et al., 2023b). However, these approaches often focus on single-step inference, limiting flexibility in complex reasoning tasks. Monte Carlo Tree Search has emerged as a powerful technique for optimizing reasoning pathways, enhancing exploration efficiency (Silver et al., 2018). It has been applied to strategic planning (Zhou et al., 2023a; Yu et al., 2023) and combined with reinforcement learning to improve LLM

reasoning without additional labeled data (Feng et al., 2023; Tian et al., 2024; Huang et al., 2024). For example, Tian et al. (2024) introduced  $\eta$ MCTS for multi-level planning, while Zhang et al. (2024) used MCTS to gather high-quality reasoning paths for training reward models. The most relevant work, rStar (Qi et al., 2024), extends MCTS by incorporating diverse reasoning operations.

**Medical RAG** enhances LLM reasoning by retrieving up-to-date knowledge for tasks such as question answering (Xiong et al., 2024a; Tian et al., 2019), classification (Li et al., 2024), and medical dialogue (Shi et al., 2024). Methods like query rewriting (Zhang et al., 2022) and multi-step retrieval (Mrini et al., 2022) further refine retrieval processes for applications in clinical decision-making and literature review (Zakka et al., 2024; Wang et al., 2023). For instance, Jeong et al. (2024) introduced SelfBioRAG, integrating retrieval with self-reflection, while i-MedRAG enables iterative refinement of retrieval queries (Xiong et al., 2024b). SeRTS (Hu et al., 2024) enhances retrieval by using MCTS with a self-rewarding mechanism, optimizing document selection via iterative exploration and fine-tuning with Proximal Policy Optimization. Unlike RARE, which integrates agentic planning with both RAG-based and non-RAG-based reasoning.

## 6 Conclusion

This work introduced **RARE (Retrieval-Augmented Reasoning Enhancement)**, a framework designed to improve the reasoning accuracy and factual reliability of large language models (LLMs) through retrieval-augmented actions and factuality scoring. RARE operates entirely as an autonomous language agent, requiring no additional training or fine-tuning of the underlying LLM. This makes the framework robust to overfitting and highly adaptable across tasks and datasets, as it relies solely on real-time retrieval and reasoning mechanisms. Experiments on medical and commonsense reasoning benchmarks demonstrate RARE’s effectiveness, RARE bridges the gap between open-source models and state-of-the-art proprietary systems, showcasing its potential as a scalable and effective solution for knowledge-intensive reasoning tasks.



## 7 Limitations

While RARE demonstrates significant improvements in reasoning accuracy and factual reliability, it has certain limitations that present opportunities for future work.

First, the framework has only been tested on open-source models like LLaMA 3.1 and not on larger proprietary models such as GPT-4. This is due to the high number of API calls required by RARE’s iterative retrieval and reasoning process, making evaluations on closed-source models prohibitively costly. However, the framework is designed to be model-agnostic and can be directly applied to proprietary models if resources permit.

Second, RARE is designed to identify a single reasoning path that leads to a correct answer but does not necessarily optimize for the best or shortest path that maximizes robustness (e.g., achieving the highest model confidence). Future work could explore designing better reward functions to prevent reward hacking and improve the selection of the most reliable reasoning paths.

Finally, RARE is currently limited to using Monte Carlo Tree Search for exploring action paths. While effective, this approach does not leverage a trained reward model to dynamically guide the search process. Future extensions could incorporate reward models or alternative optimization strategies to further enhance reasoning quality and efficiency.

These limitations highlight areas for improvement and potential research directions to make RARE more robust, generalizable, and applicable to a wider range of models and reasoning tasks.

## 8 Ethics Statement

This work aims to advance the field of Medical QA by enhancing the reasoning capabilities of language models through the RARE framework. While the results demonstrate significant improvements, several ethical considerations must be addressed to ensure responsible development and deployment:

Considering clinical applicability, RARE is designed to improve reasoning and factual reliability but is not intended to replace healthcare professionals or serve as a standalone diagnostic or treatment tool. Any integration into medical workflows must be supervised by qualified practitioners to ensure patient safety and ethical use.

Considering Bias and Fairness, Language models, including those tested with RARE, may reflect

biases present in their training data. These biases could impact the fairness and reliability of the reasoning process, particularly in sensitive medical contexts. Future work must include rigorous audits for bias and fairness to minimize potential harm.

Considering Generalizability, as RARE has been primarily evaluated in English-language, text-based general and medical domain QA tasks, its applicability to non-English-speaking contexts and multimodal scenarios remains untested. Efforts should be made to extend the framework to diverse linguistic and cultural contexts to ensure equitable access to its benefits.

Considering Societal Impacts, while RARE demonstrates the potential for improving medical reasoning tasks, its outputs should be considered supplementary to human expertise. The ethical deployment of RARE requires clear guidelines to avoid overreliance on AI and ensure that it enhances, rather than replaces, human decision-making in healthcare.

## References

- Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Bohnlshagen, Stephen Tavenor, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.
- Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.
- YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288.

Qianglong Chen, Feng Ji, Haiqing Chen, and Yin Zhang. 2020. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources. In <i>Proceedings of the 28th International Conference on Computational Linguistics</i> , pages 2583–2594.	692
Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. <i>arXiv preprint arXiv:2211.12588</i> .	693
Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Everything of thoughts: Defying the law of penrose triangle for thought generation. <i>arXiv preprint arXiv:2311.04254</i> .	694
Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. <i>arXiv preprint arXiv:2407.21783</i> .	695
Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. 2025. Competitive programming with large reasoning models. <i>arXiv preprint arXiv:2502.06807</i> .	696
Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. <i>arXiv preprint arXiv:2309.17179</i> .	697
Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. <i>Transactions of the Association for Computational Linguistics</i> , 9:346–361.	698
Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. <i>arXiv preprint arXiv:2501.12948</i> .	699
Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. <i>arXiv preprint arXiv:2305.14992</i> .	700
Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 1321–1335.	701
Jiatan Huang, Mingchen Li, Zonghai Yao, Zhichao Yang, Yongkang Xiao, Feiyun Ouyang, Xiaohan Li, Shuo Han, and Hong Yu. 2024. Ritek: A dataset for large language models complex reasoning over textual knowledge graphs. <i>arXiv preprint arXiv:2410.13987</i> .	702
Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. <i>Bioinformatics</i> , 40(Supplement_1):i119–i129.	703
Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. <i>Applied Sciences</i> , 11(14):6421.	704
Qiao Jin, Robert Leaman, and Zhiyong Lu. 2024a. Pubmed and beyond: biomedical literature search in the age of artificial intelligence. <i>EBioMedicine</i> , 100.	705
Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024b. Matching patients to clinical trials with large language models. <i>Nature Communications</i> , 15(1):9074.	706
Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In <i>Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval</i> , pages 39–48.	707
Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. <i>arXiv preprint arXiv:2210.02406</i> .	708
Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. <i>Transactions of the Association for Computational Linguistics</i> , 7:453–466.	709
Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2024. Biomedrag: A retrieval augmented large language model for biomedicine. <i>arXiv preprint arXiv:2405.00465</i> .	710
Siru Liu, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright. 2023. Using ai-generated suggestions from chatgpt to optimize clinical decision support. <i>Journal of the American Medical Informatics Association</i> , 30(7):1237–1245.	711

747	Zhiyong Lu. 2011. Pubmed and beyond: a survey	800
748	of web tools for searching biomedical literature.	801
749	<i>Database</i> , 2011:baq036.	802
750	Khalil Mrini, Harpreet Singh, Franck Dernoncourt, Se-	803
751	unghyun Yoon, Trung Bui, Walter Chang, Emilia	804
752	Farcas, and Ndapa Nakashole. 2022. Medical ques-	805
753	tion understanding and answering with knowledge	806
754	grounding and semantic self-supervision. <i>arXiv</i>	807
755	<i>preprint arXiv:2209.15301</i> .	808
756	Ankit Pal, Logesh Kumar Umapathi, and Malaikan-	809
757	nan Sankarasubbu. 2022. Medmcqa: A large-scale	810
758	multi-subject multi-choice dataset for medical do-	811
759	main question answering. In <i>Conference on health,</i>	812
760	<i>inference, and learning</i> , pages 248–260. PMLR.	813
761	Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang,	814
762	Fan Yang, and Mao Yang. 2024. Mutual reasoning	815
763	makes smaller llms stronger problem-solvers. <i>arXiv</i>	816
764	<i>preprint arXiv:2408.06195</i> .	817
765	Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hon-	818
766	gru Wang, Yuhang Guo, Tong Ruan, Jie Xu, and	819
767	Shaoting Zhang. 2024. Medical dialogue: A survey	820
768	of categories, methods, evaluation and challenges.	821
769	<i>arXiv preprint arXiv:2405.10630</i> .	822
770	David Silver, Thomas Hubert, Julian Schrittwieser, Ioan-	823
771	nis Antonoglou, Matthew Lai, Arthur Guez, Marc	824
772	Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore	825
773	Graepel, et al. 2018. A general reinforcement learn-	826
774	ing algorithm that masters chess, shogi, and go	827
775	through self-play. <i>Science</i> , 362(6419):1140–1144.	828
776	Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mah-	829
777	davi, Jason Wei, Hyung Won Chung, Nathan Scales,	830
778	Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl,	831
779	et al. 2023a. Large language models encode clinical	832
780	knowledge. <i>Nature</i> , 620(7972):172–180.	833
781	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres,	834
782	Ellery Wulczyn, Le Hou, Kevin Clark, Stephen	835
783	Pfohl, Heather Cole-Lewis, Darlene Neal, et al.	836
784	2023b. Towards expert-level medical question an-	837
785	swering with large language models. <i>arXiv preprint</i>	838
786	<i>arXiv:2305.09617</i> .	839
787	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and	840
788	Jonathan Berant. 2018. Commonsenseqa: A question	841
789	answering challenge targeting commonsense knowl-	842
790	edge. <i>arXiv preprint arXiv:1811.00937</i> .	843
791	Nandan Thakur, Nils Reimers, Andreas Rücklé, Ab-	844
792	hishek Srivastava, and Iryna Gurevych. 2021. Beir:	845
793	A heterogenous benchmark for zero-shot evalua-	846
794	tion of information retrieval models. <i>arXiv preprint</i>	847
795	<i>arXiv:2104.08663</i> .	848
796	Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian	849
797	Yu, Haitao Mi, and Dong Yu. 2024. Toward self-	850
798	improvement of llms via imagination, searching, and	851
799	criticizing. <i>arXiv preprint arXiv:2404.12253</i> .	852
	Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song.	
	2019. Chimed: A chinese medical corpus for ques-	
	tion answering. In <i>Proceedings of the 18th BioNLP</i>	
	<i>Workshop and Shared Task</i> , pages 250–260.	
	Harsh Trivedi, Niranjana Balasubramanian, Tushar Khot,	
	and Ashish Sabharwal. 2023. Interleaving retrieval	
	with chain-of-thought reasoning for knowledge-	
	intensive multi-step questions. In <i>Proceedings of the</i>	
	<i>61st Annual Meeting of the Association for Computa-</i>	
	<i>tional Linguistics (Volume 1: Long Papers)</i> , pages	
	10014–10037.	
	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le,	
	Ed Chi, Sharan Narang, Aakanksha Chowdhery, and	
	Denny Zhou. 2022. Self-consistency improves chain	
	of thought reasoning in language models. <i>arXiv</i>	
	<i>preprint arXiv:2203.11171</i> .	
	Yubo Wang, Xueguang Ma, and Wenhui Chen. 2023.	
	Augmenting black-box llms with medical textbooks	
	for clinical question answering. <i>arXiv preprint</i>	
	<i>arXiv:2309.02233</i> .	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	
	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	
	et al. 2022. Chain-of-thought prompting elicits rea-	
	soning in large language models. <i>Advances in neural</i>	
	<i>information processing systems</i> , 35:24824–24837.	
	Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu,	
	Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng,	
	Ruibo Liu, Da Huang, et al. 2024. Long-form fac-	
	tuality in large language models. <i>arXiv preprint</i>	
	<i>arXiv:2403.18802</i> .	
	Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang,	
	Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama:	
	toward building open-source language models for	
	medicine. <i>Journal of the American Medical Infor-</i>	
	<i>matics Association</i> , page ocae045.	
	Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong	
	Zhang. 2024a. Benchmarking retrieval-augmented	
	generation for medicine. <i>arXiv e-prints</i> , pages arXiv–	
	2402.	
	Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang,	
	Zhiyong Lu, and Aidong Zhang. 2024b. Im-	
	proving retrieval-augmented generation in medicine	
	with iterative follow-up questions. <i>arXiv preprint</i>	
	<i>arXiv:2408.00727</i> .	
	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Ben-	
	gio, William W Cohen, Ruslan Salakhutdinov, and	
	Christopher D Manning. 2018. Hotpotqa: A dataset	
	for diverse, explainable multi-hop question answer-	
	ing. <i>arXiv preprint arXiv:1809.09600</i> .	
	Xiao Yu, Maximillian Chen, and Zhou Yu. 2023.	
	Prompt-based monte-carlo tree search for goal-	
	oriented dialogue policy planning. <i>arXiv preprint</i>	
	<i>arXiv:2305.13660</i> .	

- Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts\*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*.
- Ming Zhang, Shuai Dou, Ziyang Wang, and Yunfang Wu. 2022. Focus-driven contrastive learning for medical question summarization. In *Proceedings of the 29th international conference on computational linguistics*, pages 6176–6186.
- Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023a. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*.
- Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2023b. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.



## A Appendix

### A.1 Implementation Details

In the MCTS framework of the Retrieval-augmented Generator, we set the number of roll-outs to 4 for LLaMA 3.2 3B and LLaMA 3.1 8B models, and 2 for the LLaMA 3.1 70B model. This configuration strikes a balance between effective inference and computational efficiency, particularly for larger models where inference costs are higher.

In the factuality scoring stage, we perform a self-scorer setup, where the Retrieval-Augmented Factuality Scorer uses the same backbone model as the generator. For instance, when the generator utilizes LLaMA 3.1 3B, the RAFS also employs the LLaMA 3.2 3B model for factuality evaluation. This ensures consistency between the generator and scorer while maintaining efficient inference. All inference processes, including factuality scoring, are parallelized to further enhance efficiency.

### A.2 Evaluation tasks

To rigorously test the reasoning capabilities of RARE, we evaluate it on a range of reasoning tasks, categorized into two main domains:

**Medical Reasoning Tasks:** We use three medical datasets that require complex, domain-specific reasoning, including:

- **MedQA-USMLE** (Jin et al., 2021): A medical question-answering dataset based on the United States Medical Licensing Examination (USMLE) questions.
- **MedMCQA** (Pal et al., 2022): A dataset consisting of multiple-choice medical questions designed to test clinical knowledge.
- **MMLU-Medical** (Singhal et al., 2023a): The medical subset of the Massive Multitask Language Understanding (MMLU) benchmark, focusing on diverse topics in the medical field.

**Commonsense Reasoning Tasks:** We evaluate RARE’s general reasoning ability on commonsense datasets. While **StrategyQA** requires complex, implicit reasoning strategies, other tasks benefit from advanced reasoning but may not require it to the same extent:

- **StrategyQA** (Geva et al., 2021): A dataset of open-domain questions that require implicit reasoning strategies.

- **CommonsenseQA** (Talmor et al., 2018): A multiple-choice question-answering dataset designed to test commonsense knowledge.

- **PIQA (Physical Interaction QA)** (Bisk et al., 2020): A dataset for physical reasoning, where models must answer questions about common physical interactions.

### A.3 Baselines

We compare the performance of RARE with several baseline reasoning methods, including:

- **Chain of Thought (CoT)** (Wei et al., 2022): A reasoning approach that generates explanations step-by-step, aiming for more coherent answers.
- **Self-Consistency** (Wang et al., 2022): A method that uses majority voting among multiple reasoning paths to increase response accuracy.
- **rStar** (Qi et al., 2024): A framework that extends MCTS with a diverse set of reasoning actions, improving reasoning accuracy.
- **MedRAG** (Xiong et al., 2024a): A retrieval-augmented framework designed for medical question answering, integrating knowledge retrieval from domain-specific corpora.
- **i-MedRAG** (Xiong et al., 2024b): An iterative retrieval-based medical reasoning model that dynamically refines queries to improve response accuracy in medical QA tasks.

### A.4 Retrieval Model and Corpus

Corpus	#Docs	#Snippets	Avg. L	Domain
PubMed	23.9M	23.9M	296	Biomed.
StatPearls	9.3k	301.2k	119	Clinics
Textbooks	18	125.8k	182	Medicine
Wikipedia	6.5M	29.9M	162	General
MedCorp	30.4M	54.2M	221	Mixed

Table 5: Statistics of the retrieval corpus used in our experiments. #Docs refers to the number of documents, #Snippets represents extracted text units, and Avg. L indicates the average snippet length.

For information retrieval, we leverage the **Med-Corp** (Xiong et al., 2024a) corpus, a curated collection of high-quality and domain-specific resources

that serve as knowledge bases for medical and general question answering. The **MedCorp** corpus consists of (table 5):

- **PubMed**<sup>2</sup>: A widely used biomedical literature database containing over 36 million articles (Lu, 2011; Jin et al., 2024a). For our retrieval tasks, we utilize a subset of **23.9 million** articles with valid titles and abstracts, similar to the MedRAG setup (Xiong et al., 2024a).
- **StatPearls**<sup>3</sup>: A clinical decision support resource with publicly available medical articles hosted on **NCBI Bookshelf**<sup>4</sup>. The corpus includes **9,330** peer-reviewed StatPearls articles, structured into hierarchical snippets where each paragraph is treated as a retrieval unit, with corresponding hierarchical headings as metadata.
- **Medical Textbooks**<sup>5</sup>: A collection of **18 widely used medical textbooks** (Jin et al., 2021) that are commonly referenced for foundational medical knowledge and United States Medical Licensing Examination (USMLE) preparation. The textbook corpus is segmented into passages of up to **1,000 characters** using the RecursiveCharacterTextSplitter from LangChain<sup>6</sup>.
- **Wikipedia**: A large-scale, general-domain encyclopedia frequently used in information retrieval tasks (Thakur et al., 2021). We incorporate a **processed version** of Wikipedia from Huggingface<sup>7</sup> and apply text chunking techniques to facilitate passage-level retrieval.

To retrieve relevant passages from **MedCorp**, we employ the **ColBERT retrieval model** (Khattab and Zaharia, 2020), an efficient neural retrieval framework optimized for **passage ranking**. ColBERT’s **late interaction mechanism** enables fine-grained token-level relevance matching, allowing our retrieval-augmented generator to identify and incorporate **contextually relevant** medical knowledge. This approach enhances the factual accuracy and reliability of generated responses while efficiently handling large-scale biomedical corpora.

<sup>2</sup><https://pubmed.ncbi.nlm.nih.gov/>

<sup>3</sup><https://www.statpearls.com/>

<sup>4</sup><https://www.ncbi.nlm.nih.gov/books/NBK430685/>

<sup>5</sup><https://github.com/jind11/MedQA>

<sup>6</sup><https://www.langchain.com/>

<sup>7</sup><https://huggingface.co/datasets/wikipedia>

## A.5 Monte Carlo Tree Search (MCTS)

Monte Carlo Tree Search (MCTS) is a decision-making algorithm widely used in complex decision processes, such as games, by building a search tree and simulating outcomes to estimate the value of potential actions. MCTS operates through four main phases (Browne et al., 2012).

**Selection:** Starting from the root node, the algorithm traverses through child nodes based on strategies like the Upper Confidence Bound applied on Trees (UCT), which balances exploration and exploitation, continuing until a leaf node is reached.

**Expansion:** At the leaf node, if it does not represent a terminal state, one or more feasible child nodes are added to represent potential future actions.

**Simulation (Evaluation):** From one of the newly added nodes (typically selected randomly), random simulations (or "rollouts") are performed by selecting actions randomly until reaching a terminal state, thereby estimating the node’s potential.

**Backpropagation:** After simulation, the results (win, loss, or draw) are propagated back through the traversed nodes, updating the statistical data (e.g., rewards or visit counts) to guide future decision-making.

By iterating through these phases, MCTS incrementally builds a decision tree, enabling optimal strategy refinement in scenarios where direct calculation of the best strategy is infeasible due to a vast state space. A crucial component of MCTS is the Upper Confidence Bound applied on Trees (UCT) algorithm, used during the selection phase to balance exploration and exploitation. The UCT formula for choosing actions is defined as follows:

$$UCT_j = \bar{X}_j + C \sqrt{\frac{2 \ln N}{N_j}} \quad (1)$$

where  $\bar{X}_j$  is the average reward of action  $j$ ,  $N$  is the total visit count of the parent node,  $N_j$  is the visit count of node  $j$ , and  $C$  is a constant that controls the balance between exploration and exploitation.

rStar enhanced MCTS-based exploration of candidate solutions. Specifically, rStar leverages a reward mechanism to guide tree expansion. Each node  $s$  generated under action  $a$  has a reward value  $Q(s, a)$ . Unexplored nodes are initialized with  $Q(s_i, a_i) = 0$ , leading to random tree expansions initially. Upon reaching a terminal node  $s_d$ , a reward score  $Q(s_d, a_d)$  is computed based on

whether the trajectory reaches the correct answer and is then back-propagated to each intermediate node in the trajectory  $t = x \oplus s_1 \oplus s_2 \oplus \dots \oplus s_d$ . For each intermediate node  $s_i$ , the reward is updated as  $Q(s_i, a_i) = Q(s_i, a_i) + Q(s_d, a_d)$ , with  $Q(s_d, a_d)$  determined by the likelihood or confidence from self-consistency majority voting. This reward propagation directs the search toward paths with higher probabilities of correctness, enhancing decision-making within the reasoning process.

#### A.6 Evaluation of Reasoning Candidates with RAFS

Figure 7 demonstrates how the Retrieval-Augmented Factuality Scorer (RAFS) evaluates reasoning candidates for a medical question. The example involves diagnosing and treating a patient presenting with symptoms consistent with seasonal allergic conjunctivitis. RAFS assigns factuality scores to each candidate reasoning path based on their alignment with retrieved evidence.

**Candidate Reasoning 1** accurately identifies the condition and treatment, correctly selecting ketotifen eye drops (Answer B) with a factuality score of 1.0. **Candidate Reasoning 2** suggests fluorometholone eye drops (Answer D) but includes conflicting reasoning and inaccuracies, resulting in a lower factuality score of 0.625. **Candidate Reasoning 3** incorrectly recommends warm compresses (Answer C) as the most appropriate treatment, reflecting limited factual alignment and receiving a score of 0.6.

This example highlights RAFS’s ability to differentiate between reasoning paths based on factual correctness, ensuring that the most reliable and evidence-supported answer is prioritized.

#### A.7 Prompt of Different Actions in MCTS

### RAFS Examples

**Question:** A 35-year-old man comes to the physician because of itchy, watery eyes for the past week. He has also been sneezing multiple times a day during this period. He had a similar episode 1 year ago around springtime. He has iron deficiency anemia and ankylosing spondylitis. Current medications include ferrous sulfate, artificial tear drops, and indomethacin. He works as an elementary school teacher. His vital signs are within normal limits. Visual acuity is 20/20 without correction. Physical examination shows bilateral conjunctival injection with watery discharge. The pupils are 3 mm, equal, and reactive to light. Examination of the anterior chamber of the eye is unremarkable. Which of the following is the most appropriate treatment? A: Erythromycin ointment, B: Ketotifen eye drops, C: Warm compresses, D: Fluorometholone eye drops

**Candidate Reasoning 1:** Let's think step by step. Based on the symptoms described - itchy, watery eyes, sneezing, bilateral conjunctival injection, and watery discharge - along with the history of a similar episode around springtime, it's clear that the patient is experiencing allergic conjunctivitis, likely triggered by seasonal allergens like pollen given his profession as an elementary school teacher. For symptomatic relief of allergic conjunctivitis, antihistamine eyedrops or mast cell stabilizers can be effective. Among the options provided, ketotifen eye drops act as a mast cell stabilizer and would be appropriate for treating allergic conjunctivitis. The answer is B: Ketotifen eye drops. ✓ **The factuality score is 1.0**

**Candidate Reasoning 2:** Based on the patient's symptoms of itchy, watery eyes, sneezing, and physical examination findings of bilateral conjunctival injection with watery discharge, the diagnosis is allergic conjunctivitis. Considering the timing of the episode and the patient's profession, the most likely causative agent is seasonal pollen. Treatment should focus on reducing symptoms rather than eliminating the allergen source. Fluorometholone eye drops are corticosteroid drops that reduce inflammation and can provide relief from itching and redness. Other options like erythromycin or ketotifen may have some effect but would be less effective in addressing this presentation. Warm compresses might help with discharge, but again, would be less effective compared to the impact of corticosteroids. Therefore, the most appropriate treatment is fluorometholone eye drops. The answer is D: Fluorometholone eye drops. ✗ **The factuality score is 0.625**

**Candidate Reasoning 3:** Given the patient's symptoms of itchy, watery eyes, sneezing, and conjunctival injection, along with a history of similar episodes around springtime, this case is most consistent with seasonal allergic conjunctivitis. The best treatment for mild allergic conjunctivitis involves avoiding triggers when possible, using lubricating artificial tears regularly, and applying a cold compress. Warm compresses are often recommended to help loosen crusts and debris, improving comfort. Therefore, warm compresses would be the most appropriate treatment for this patient. The answer is C: Warm compresses. ✗ **The factuality score is 0.6**

Figure 7: Evaluation of reasoning candidates by the Retrieval-Augmented Factuality Scorer (RAFS). Each candidate provides a reasoning path for a given question, with RAFS assigning a factuality score based on alignment with retrieved evidence.



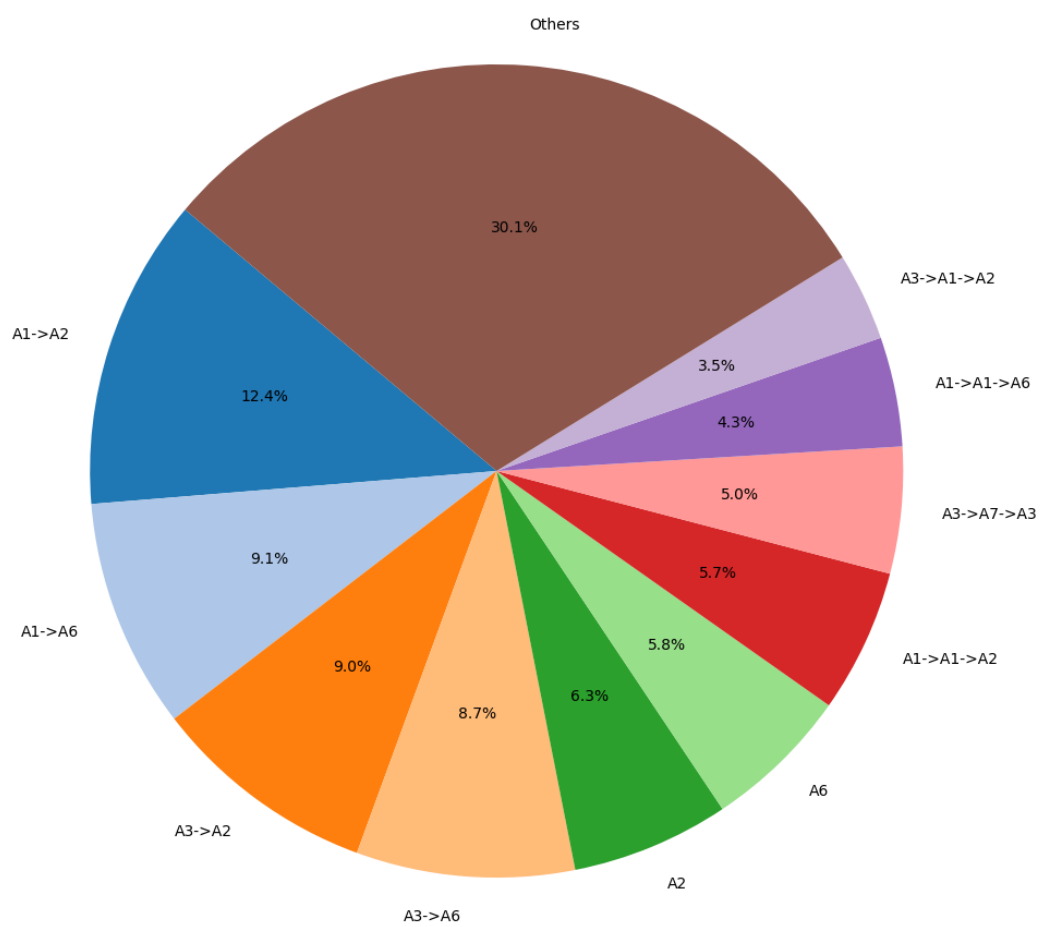


Figure 8: Top 10 common paths that lead to correct answer (MedQA)

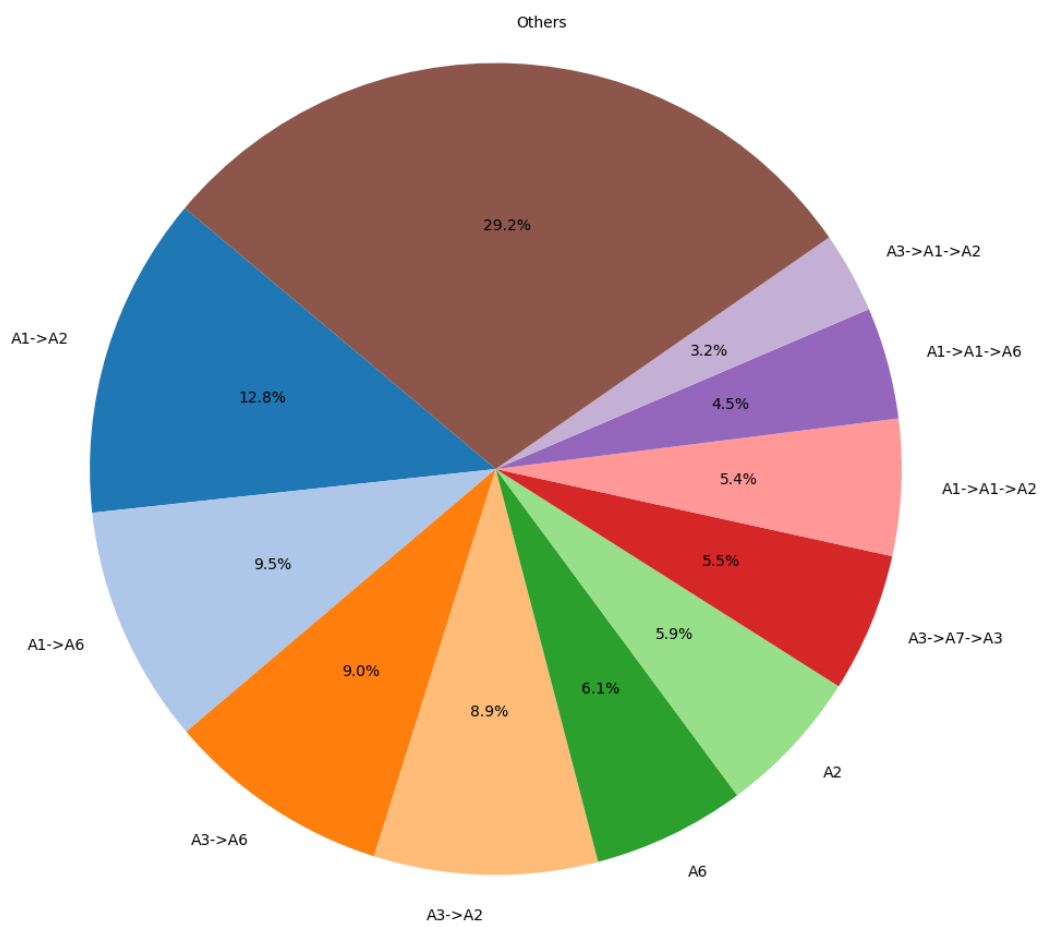


Figure 9: Top 10 common paths that lead to correct answer (StrategyQA)

### Multi-step reasoning in Medical QA

**Question:** A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient? A: Ampicillin, B: Ceftriaxone, C:Ciprofloxacin, D: Doxycycline, E: Nitrofurantoin

**Step 1: Identify key clinical features** A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination (dysuria), worsening symptoms over one day, and no systemic signs of infection (normal temperature, blood pressure, and oxygen saturation). Physical exam shows no costovertebral angle tenderness, suggesting the absence of pyelonephritis.

**Step 2: Determine the likely diagnosis** Dysuria without systemic signs in a pregnant woman is most consistent with acute cystitis, a lower urinary tract infection.

**Step 3: Evaluate treatment options for cystitis in pregnancy** Nitrofurantoin: Safe and effective during the second trimester, commonly used for uncomplicated cystitis. Ampicillin: Can be used but has high resistance rates among common uropathogens like E. coli. Ceftriaxone: Typically used for severe infections like pyelonephritis and administered intravenously, not necessary here. Ciprofloxacin: A fluoroquinolone, contraindicated in pregnancy due to potential cartilage damage to the fetus. Doxycycline: A tetracycline, contraindicated in pregnancy due to risks of fetal bone growth inhibition and teeth discoloration.

**Step 4: Select the best treatment** Nitrofurantoin is the safest and most effective option for this patient in the second trimester.

**Step 5: Conclusion** The best treatment for this patient is nitrofurantoin. The answer is E: Nitrofurantoin.

Figure 10: An example of a medical question that requires multi-step reasoning to answer.

Action 1. Propose an one-step thought.

### Instruction: A 46-year-old woman comes to the physician because of a 2-week history of diplopia and ocular pain when reading the newspaper. She also has a 3-month history of amenorrhea, hot flashes, and increased sweating. She reports that she has been overweight all her adult life and is happy to have lost 6.8 kg (15 lb) of weight in the past 2 months. Her pulse is 110/min, and blood pressure is 148/98 mm Hg. Physical examination shows moist palms and a nontender thyroid gland that is enlarged to two times its normal size. Ophthalmologic examination shows prominence of the globes of the eyes, bilateral lid retraction, conjunctival injection, and an inability to converge the eyes. There is no pain on movement of the extraocular muscles. Visual acuity is 20/20 bilaterally. Neurologic examination shows a fine resting tremor of the hands. Deep tendon reflexes are 3+ with a shortened relaxation phase. Which of the following is the most likely cause of this patient's ocular complaints? A: Granulomatous inflammation of the cavernous sinus, B: Abnormal communication between the cavernous sinus and the internal carotid artery, C: Glycosaminoglycan accumulation in the orbit, D: Bacterial infection of the orbital contents, E: Sympathetic hyperactivity of levator palpebrae superioris

### Response: Let's think step by step.

Step 1: Identify key clinical features: A 46-year-old woman presents with diplopia, ocular pain, weight loss, amenorrhea, and hyperthyroid signs like tachycardia, enlarged thyroid gland, and hyperreflexia.

Step 2: Recognize underlying condition: These findings suggest Graves' disease with thyroid-associated orbitopathy (TAO).

Step 3: Determine cause of ocular symptoms: Prominent globes, lid retraction, and conjunctival injection are caused by immune-mediated glycosaminoglycan accumulation in the orbit, characteristic of TAO.

Step 4: Eliminate other options: Conditions like cavernous sinus inflammation, carotid-cavernous fistula, and bacterial infection do not fit the patient's presentation.

Step 5: The most likely cause of the ocular complaints is glycosaminoglycan accumulation in the orbit.

Step 6: Conclusion: The answer is C: Glycosaminoglycan accumulation in the orbit.

### Instruction: {user question}

### Response: Let's think step by step.



Action 2: Propose the remaining thought steps / Action 4: Answer the sub-question again.

### Instruction: A 1-year-old boy presents to the emergency department with weakness and a change in his behavior. His parents state that they first noticed the change in his behavior this morning and it has been getting worse. They noticed the patient was initially weak in his upper body and arms, but now he won't move his legs with as much strength or vigor as he used to. Physical exam is notable for bilateral ptosis with a sluggish pupillary response, a very weak sucking and gag reflex, and shallow respirations. The patient is currently drooling and his diaper is dry. The parents state he has not had a bowel movement in over 1 day. Which of the following is the pathophysiology of this patient's condition? A: Antibodies against postsynaptic nicotinic cholinergic ion channels, B: Autoantibodies against the presynaptic voltage-gated calcium channels, C: Autoimmune demyelination of peripheral nerves, D: Blockade of presynaptic acetylcholine release at the neuromuscular junction, E: Lower motor neuron destruction in the anterior horn

### Response: Let's think step by step. The infant has rapidly progressing weakness starting from the upper body and moving downward, bilateral ptosis, sluggish pupils, weak sucking and gag reflexes, shallow respirations, drooling, constipation (dry diaper and no bowel movement), and decreased muscle strength. These symptoms are consistent with infant botulism, caused by ingestion of *Clostridium botulinum* spores (often from honey) that produce botulinum toxin. The toxin inhibits the release of acetylcholine at the presynaptic terminals of the neuromuscular junction by cleaving SNARE proteins necessary for vesicle fusion. This blockade leads to flaccid paralysis and autonomic dysfunction. Therefore, the pathophysiology is blockade of presynaptic acetylcholine release at the neuromuscular junction. The answer is D: Blockade of presynaptic acetylcholine release at the neuromuscular junction.

### Instruction: {user question}

### Response:

Action 3: Propose next sub-question along with its answer.

Given a question, please decompose it into sub-questions. For each sub-question, please answer it in a complete sentence, ending with "The answer is ". When the original question is answerable, please start the subquestion with "Now we can answer the question: <original question>".

Question 1: A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient? A: Ampicillin, B: Ceftriaxone, C: Ciprofloxacin, D: Doxycycline, E: Nitrofurantoin

Question 1.1: What is the most likely diagnosis for this patient?

Answer 1.1: The patient is a 23-year-old pregnant woman at 22 weeks gestation presenting with burning upon urination, known as dysuria. She has no fever, normal vital signs, and no costovertebral angle tenderness. This suggests that she is experiencing an acute uncomplicated urinary tract infection (UTI), specifically cystitis.

Question 1.2: Which antibiotics are safe and effective for treating cystitis during pregnancy?

Answer 1.2: During pregnancy, certain antibiotics are preferred due to their safety profiles for both the mother and the fetus. Nitrofurantoin is commonly used and considered safe during the second trimester for treating cystitis. Ampicillin may be less effective due to resistance. Ceftriaxone is typically reserved for more severe infections like pyelonephritis and is administered intravenously. Ciprofloxacin and doxycycline are contraindicated in pregnancy because they can harm fetal development.

Question 1.3: Now we can answer the question: Which of the following is the best treatment for this patient?

Answer 1.3: The best treatment for this pregnant patient with acute cystitis is Nitrofurantoin, as it is safe during the second trimester and effective against common urinary pathogens. Therefore, the answer is E: Nitrofurantoin.

Question 2: {user question}

#### Action 5: Rephrase the question/sub-question.

You are an AI assistant to help me rephrase questions by splitting the question context into conditions. In your rephrased question, remember to fully express the information in the original question.

Original Question: A 62-year-old woman comes to the physician because of coughing and fatigue during the past 2 years. In the morning, the cough is productive of white phlegm. She becomes short of breath walking up a flight of stairs. She has hypertension and hyperlipidemia. She has recently retired from working as a nurse at a homeless shelter. She has smoked 1 pack of cigarettes daily for 40 years. Current medications include ramipril and fenofibrate. Her temperature is 36.5°C (97.7°F), respirations are 24/min, pulse is 85/min, and blood pressure is 140/90 mm Hg. Scattered wheezing and rhonchi are heard throughout both lung fields. There are no murmurs, rubs, or gallops but heart sounds are distant. Which of the following is the most likely underlying cause of this patient's symptoms? A: Chronic decrease in pulmonary compliance, B: Local accumulation of kinins, C: Mycobacterial invasion of pulmonary parenchyma, D: Progressive obstruction of expiratory airflow, E: Incremental loss of functional residual capacity

Rephrased Question: A 62-year-old woman comes to the physician due to a 2-year history of coughing and fatigue. She has a morning cough productive of white phlegm and becomes short of breath when climbing stairs. She has hypertension and hyperlipidemia and recently retired from nursing at a homeless shelter. She has a 40-pack-year smoking history. Her medications include ramipril and fenofibrate. Vital signs are temperature 36.5°C (97.7°F), respirations 24/min, pulse 85/min, and blood pressure 140/90 mmHg. Physical examination reveals scattered wheezing and rhonchi throughout both lung fields. Heart sounds are distant with no murmurs, rubs, or gallops. Which of the following is the most likely underlying cause of this patient's symptoms? A: Chronic decrease in pulmonary compliance, B: Local accumulation of kinins, C: Mycobacterial invasion of pulmonary parenchyma, D: Progressive obstruction of expiratory airflow, E: Incremental loss of functional residual capacity

Original question: {user question}

Rephrased Question:

#### Action 6: Generate multiple search queries and retrieve information.

Given a question, generate a search query that would help gather information to answer it. Your goal is to formulate a query that will retrieve useful evidence or additional details that contribute to answering the question. The query should aim to obtain new information and be specific enough to ensure that the search results are relevant and helpful.

Question 1: A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient? A: Ampicillin, B: Ceftriaxone, C: Ciprofloxacin, D: Doxycycline, E: Nitrofurantoin

Query 1.1: Common causes and management of dysuria in pregnant women without systemic symptoms.

Document 1.1: Dysuria in pregnant women without systemic symptoms like fever or flank pain is often indicative of acute cystitis, a lower urinary tract infection typically caused by bacteria such as *Escherichia coli*. Pregnancy increases the risk of urinary tract infections due to hormonal changes that relax ureteral smooth muscle and mechanical compression from the enlarging uterus, leading to urinary stasis. Management involves confirming the diagnosis with a urine culture and initiating antibiotic therapy that is safe for use during pregnancy.

Query 1.2: Antibiotics that are safe and effective for treating cystitis during pregnancy.

Document 1.2: Safe and effective antibiotics for treating cystitis in pregnancy include nitrofurantoin and certain beta-lactam antibiotics like ampicillin and cephalexin. Nitrofurantoin is commonly used during the second trimester and is effective against common urinary pathogens. Ampicillin can be used but may have increased resistance rates. Ceftriaxone is a third-generation cephalosporin administered intravenously and is typically reserved for more severe infections like pyelonephritis. Antibiotics such as doxycycline (a tetracycline) and ciprofloxacin (a fluoroquinolone) are contraindicated in pregnancy due to potential risks to fetal development.

Query 1.3: Antibiotics contraindicated during pregnancy and their associated risks.

Document 1.3: Certain antibiotics are contraindicated during pregnancy due to teratogenic effects or adverse outcomes for the fetus. Doxycycline, a tetracycline antibiotic, is contraindicated because it can inhibit bone growth and cause permanent tooth discoloration in the developing fetus. Ciprofloxacin, a fluoroquinolone, is avoided as it has been associated with cartilage damage in animal studies and potential musculoskeletal risks in humans. These medications should not be used to treat infections in pregnant patients.

Question 2: {user question}

Action 7: Re-answer the question/sub-question with retrieved information.

### Instruction: A 1-year-old boy presents to the emergency department with weakness and a change in his behavior. His parents state that they first noticed the change in his behavior this morning and it has been getting worse. They noticed the patient was initially weak in his upper body and arms, but now he won't move his legs with as much strength or vigor as he used to. Physical exam is notable for bilateral ptosis with a sluggish pupillary response, a very weak sucking and gag reflex, and shallow respirations. The patient is currently drooling and his diaper is dry. The parents state he has not had a bowel movement in over 1 day. Which of the following is the pathophysiology of this patient's condition? A: Antibodies against postsynaptic nicotinic cholinergic ion channels, B: Autoantibodies against the presynaptic voltage-gated calcium channels, C: Autoimmune demyelination of peripheral nerves, D: Blockade of presynaptic acetylcholine release at the neuromuscular junction, E: Lower motor neuron destruction in the anterior horn

### Relevant Documents:

Infant Botulism: Caused by ingestion of *Clostridium botulinum* spores (common sources include honey and soil). Spores germinate in the infant's gut, producing botulinum toxin.

Botulinum Toxin Mechanism: Blocks the release of acetylcholine at presynaptic neuromuscular junctions by cleaving SNARE proteins. Results in flaccid paralysis and autonomic dysfunction.

Clinical Features: Symmetric descending paralysis starting with cranial nerves (ptosis, poor suck, weak gag reflex).

Autonomic symptoms: constipation, dry mouth, drooling. Hypotonia and respiratory difficulties.

### Response: Let's think step by step. The infant presents with rapidly progressing weakness starting in the upper body and moving downward, bilateral ptosis with sluggish pupillary response, weak sucking and gag reflexes, shallow respirations, drooling, and constipation (dry diaper, no bowel movement). These symptoms are characteristic of infant botulism, caused by ingestion of *Clostridium botulinum* spores that produce botulinum toxin. The toxin inhibits the release of acetylcholine at the presynaptic terminals of the neuromuscular junction by cleaving SNARE proteins necessary for neurotransmitter release. This blockade leads to flaccid paralysis and autonomic dysfunction. Therefore, the pathophysiology is blockade of presynaptic acetylcholine release at the neuromuscular junction. The answer is D: Blockade of presynaptic acetylcholine release at the neuromuscular junction.

### Instruction: {user question}

### Relevant Documents: {relevant documents}

### Response: