Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25665

Logical Reasoning Capabilities of Large Language Models: A Comparative Evaluation on GPQA Dataset

Boqiao Wan

The University of Melbourne, Parkville, Australia
boqiaow@gmail.com

Abstract. Recent advancements in Large Language Models (LLMs) have significantly enhanced their logical reasoning capabilities, presenting new opportunities for applications in scientific reasoning tasks. This study systematically evaluates and compares the logical reasoning performance of five prominent LLMs—GPT-4o, OpenAI-o3, OpenAI-o3 pro, DeepSeek V3, and DeepSeek R1—using the GPQA dataset, a standardized collection of science-related multiple-choice questions spanning biology, chemistry, and physics. Three dimensions were analyzed: overall accuracy, response time, and performance across difficulty levels.Results indicate that all tested models outperformed human experts in overall accuracy. Particularly, models utilizing deep-thinking (Chain-of-Thought) mechanisms consistently surpassed those without, underscoring the effectiveness of advanced reasoning strategies. Domain-specific analyses revealed superior performance on structured computational tasks in physics but relatively weaker performance in chemistry questions involving complex organic inference. Notably, increased processing time (as observed with OpenAI-o3 pro) did not proportionally enhance accuracy. Detailed analysis suggests this discrepancy was due not to resource constraints but to inefficiencies in the model's reasoning or exploratory pathways, as it frequently expended additional time retrieving descriptive but non-essential background information. Further investigation into these reasoning bottlenecks is necessary to better understand and overcome these limitations, providing valuable insights for future research and model improvements.

Keywords: Large Language Models, logical reasoning, Chain-of-Thought, model evaluation, Artificial Intelligence

1. Introduction

In recent years, artificial intelligence (AI) technologies have advanced rapidly, especially in the fields of natural language processing (NLP) and machine learning (ML), achieving significant breakthroughs [1]. Among these developments, Large Language Models (LLMs) represent a pivotal technological innovation, becoming increasingly integrated into various fields such as education, scientific research, and knowledge assessment [2]. These models not only excel at conventional information retrieval and language generation tasks but also demonstrate growing potential in addressing complex logical reasoning and problem-solving challenges [3].

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

Currently, an effective approach for evaluating the logical reasoning abilities of AI models involves utilizing standardized and structured academic tests [4]. This method facilitates quantitative analysis and intuitively reflects the models' performance in logical analysis, reasoning judgment, and comprehensive problem-solving [5]. Although considerable research has explored the capabilities of AI models, most studies focus primarily on evaluating singular capabilities such as language comprehension and text generation [2, 6, 7]. Comparatively, systematic and comprehensive research involving horizontal comparisons of logical reasoning abilities across multiple mainstream models remains notably limited, highlighting a significant research gap in this field.

Based on the research context outlined above, this study aims to systematically analyze the logical reasoning and comprehensive problem-solving performance of leading Large Language Models (LLMs), such as OpenAI's GPT series and the DeepSeek series, through cross-model and cross-domain comparisons. Specifically, this research addresses the following key questions:

- Are there significant differences in logical reasoning capabilities among various LLMs when solving science-related multiple-choice questions? How do these differences compare to the reasoning performance of human experts?
- Do these models exhibit stable and consistent logical reasoning performance when addressing questions across different scientific domains?
- Does activating internal deep reasoning functions (such as the chain-of-thought prompting technique) significantly enhance logical reasoning performance? Moreover, is the additional computational cost incurred by these advanced reasoning methods justified and acceptable for practical use?

By exploring these research questions, this study aims to provide empirical insights and theoretical considerations to support the future development and optimization of Large Language Models.

2. Related work

2.1. Recent developments in AI and large language models

In the continuous iteration and evolution of AI models, enhancing logical reasoning capabilities has progressively become a central focus of major research and development institutions. For example, OpenAI introduced the OpenAI O1 model in 2024, specifically emphasizing advanced logical reasoning. Leveraging reinforcement learning methodologies, this model is designed to generate detailed and coherent internal reasoning chains, significantly improving its performance in scientific inference, mathematical problem-solving, and code analysis tasks [8]. Subsequently, in 2025, OpenAI further advanced this technology with the release of the more powerful OpenAI o3 series, which not only expanded significantly in terms of parameter scale and context-processing capabilities but also integrated advanced analytical tools and a Python programming environment. These enhancements allowed o3 to set new benchmarks in prominent evaluations such as Codeforces, SWE-bench, and MMMU, demonstrating industry-leading logical reasoning performance [9].

Simultaneously, in 2025, DeepSeek built upon its foundational DeepSeek V3 model to launch the DeepSeek R1 model, which specifically aims to enhance logical reasoning through large-scale reinforcement learning and a "deep-thinking" mechanism. DeepSeek R1 demonstrated substantial improvements in logical inference performance, notably achieving an accuracy increase from 70%

to 87.5% on challenging tests such as the 2025 American Invitational Mathematics Examination (AIME), underscoring the model's robust capacity for complex analytical and reasoning tasks [10].

These recent advancements, highlighted by the impressive performance of the three aforementioned models in authoritative evaluations, underscore the growing importance of logical reasoning as a central competitive benchmark and pivotal area of future AI development [3, 11].

2.2. Evolution of logical reasoning evaluation and its scientific applications

In recent years, with the continuous advancement in logical reasoning capabilities of AI models and the expansion of their application scenarios, corresponding datasets and evaluation standards have also evolved rapidly. A number of novel benchmarks specifically designed to assess complex logical reasoning have emerged, including datasets such as GSM8K for evaluating mathematical reasoning and multi-step computational tasks [12], the more challenging MATH dataset focused on advanced mathematical problem-solving [13], the LogiQA dataset emphasizing logical reasoning within linguistic analysis tasks [14], and comprehensive benchmarks like Big-Bench, which cover tasks ranging from common-sense reasoning and logical inference to complex cross-domain problem-solving [15].

Compared to earlier datasets, these recent benchmarks significantly improve in terms of scale, difficulty, and task diversity, providing a more effective evaluation of AI models' real-world reasoning and generalization abilities [16]. Meanwhile, evaluation standards have progressed from relying solely on accuracy to incorporating more sophisticated dimensions, such as coherence of reasoning chains (Chain-of-Thought, CoT), interpretability of explanations, and cross-modal and cross-task reasoning capabilities [17-19]. Similarly, benchmarks targeting program reasoning and code analysis, such as SWE-bench and Codeforces, impose higher standards on the rigor and precision of logical inference processes [9].

Benefiting from these enhancements in logical reasoning, contemporary AI models have gradually become widely applied in scientific fields including mathematics, physics, chemistry, and biology, demonstrating notably strong performance in mathematical Olympiads, international standardized tests, and multi-step scientific problem-solving tasks [9, 13, 20]. Additionally, these models have facilitated the development of automated educational support and knowledge assessment systems and effectively assisted researchers in formulating scientific hypotheses, exploring theoretical concepts, and analyzing experimental data, increasingly serving as indispensable tools in scientific research and educational contexts [5, 11]. However, current AI models still face certain challenges when handling highly complex and structured reasoning tasks. Therefore, further enhancing the depth and generalization capabilities of logical reasoning in AI models remains an important direction for future research.

3. Methodology

This study aims to evaluate and compare the logical reasoning capabilities of two prominent Large Language Model (LLM) series—namely, the GPT series (GPT-4o, OpenAI-o3, OpenAI-o3 Pro) developed by OpenAI, and the DeepSeek series (DeepSeek V3, DeepSeek R1). The GPQA dataset is employed for this evaluation, specifically targeting three scientific disciplines: physics, chemistry, and biology, comprising a set of multiple-choice questions designed explicitly to assess logical inference and analytical reasoning skills. The primary evaluation metrics include accuracy across different scientific domains and difficulty levels, as well as the average response time specifically recorded for models employing deep-thinking mechanisms. Additionally, for a more comprehensive understanding of model performance, the results of these models are compared against the responses of human experts. This comparative analysis aims to highlight both the strengths and limitations of current mainstream AI models in logical reasoning tasks and provide empirical data for reference in subsequent research.

3.1. Selection of AI system

This study selected five Large Language Models (LLMs) for analysis, each chosen with specific considerations to comprehensively evaluate their logical reasoning capabilities:

- GPT-4o: The latest iteration in OpenAI’s GPT series, representing state-of-the-art language processing and logical reasoning capabilities. Including GPT-4o provides a robust baseline to analyze the evolutionary trends and reasoning performance within the GPT model lineage.
- OpenAI-o3: Optimized specifically for deeper logical inference and analytical reasoning tasks. Its inclusion enables targeted evaluation of the practical effectiveness of enhanced logical frameworks.
- OpenAI-o3 Pro: An advanced variant extending reasoning processing time beyond the o3 model, theoretically enabling more precise and nuanced logical outputs. This study assesses whether the extended processing duration significantly improves performance on challenging logical reasoning tasks.
- DeepSeek V3: Developed by DeepSeek, this foundational model represents the baseline of logical reasoning capability within China’s AI landscape. Its analysis enables direct comparison with subsequent enhanced models to illustrate iterative improvements.
- DeepSeek R1: Built upon the DeepSeek V3 architecture, this model incorporates specialized advanced reasoning algorithms and deeper inference strategies, specifically targeting complex logical reasoning tasks. The study aims to evaluate performance improvements resulting from these specialized optimization approaches.

By systematically analyzing these representative models from the two prominent model series (OpenAI’s GPT and DeepSeek), this study provides detailed insights into their comparative performance and developmental trajectories on complex logical reasoning tasks.

3.2. Selection of dataset

In this study, the GPQA dataset is employed as the primary test set, which is specifically designed for the scientific domain and covers a wide range of logical reasoning and analytical inference questions. This makes it particularly suitable for evaluating the logical reasoning and problem-solving capabilities of AI models in science-related fields. Since the GPQA dataset exclusively includes questions from three scientific disciplines—physics, chemistry, and biology—thus, the evaluation in this study focuses on these specific areas.

Specifically, 135 questions were selected from each of these three disciplines. The questions are categorized into four difficulty levels: "Easy undergraduate level," "Hard undergraduate level," "Hard graduate level," and "Post-graduate level or harder." Due to the extremely limited number of "Easy undergraduate level" questions (fewer than two questions per discipline), this category was excluded from the formal analysis. For the remaining three levels, 20 questions each were proportionally selected from "Hard undergraduate level" and "Hard graduate level," and 5 questions from "Post-graduate level or harder," ensuring that the difficulty distribution of selected questions was both balanced and adequately representative.

3.3. Evaluation strategy

In this study, the following evaluation metrics are employed to comprehensively assess the logical reasoning capabilities of the models:

• Total Accuracy:Each correctly answered question from the GPQA dataset is awarded one point. The total accuracy for each model is thus calculated as the sum of points obtained across all selected questions. This metric directly reflects the overall logical reasoning accuracy of the models.

• Accuracy by Difficulty Level:To gain deeper insights into the models' performance at varying difficulty levels, accuracy is calculated separately for the three defined difficulty categories: "Hard undergraduate level," "Hard graduate level," and "Post-graduate level or harder." This stratified scoring approach allows us to evaluate differences in reasoning proficiency across questions of varying complexity.

• Accuracy by Scientific Domain:To further examine the logical reasoning performance of models within distinct scientific disciplines, accuracy is computed individually for the physics, chemistry, and biology domains. This metric helps to reveal differences in the models' domain-specific knowledge and reasoning capabilities.

• Average Response Time (for Deep-Thinking Models only):For models employing a deep-thinking mechanism (such as OpenAI-o3 Pro and DeepSeek R1), additional processing time is utilized to optimize reasoning processes. Therefore, the average time taken by these models to respond to each question is recorded, aiming to evaluate whether extending reasoning time significantly enhances their performance on complex logical reasoning tasks.

4. Experiment results

Table 1 summarizes the logical reasoning performance of the selected Large Language Models (namely, GPT-4o, OpenAI-o3, OpenAI-o3 Pro, DeepSeek V3, and DeepSeek R1) evaluated on the GPQA dataset across three scientific domains (biology, chemistry, and physics). Performance evaluation included multiple dimensions, such as total accuracy, accuracy across different difficulty levels (Hard Undergraduate, Hard Graduate, and Post-graduate or harder), and average response time for models employing deep-thinking mechanisms (OpenAI-o3, OpenAI-o3 Pro, and DeepSeek R1). Additionally, model performance was compared against human experts as a reference benchmark.

For accuracy metrics, values were rounded to one decimal place. Due to the larger numerical values recorded in seconds, average response times were rounded to the nearest integer. Overall, the results reveal notable variations among models and across scientific domains. Detailed analyses of these evaluation metrics will be provided in the subsequent sections.

Table 1. Model performance comparison across scientific domain

Model	Scientific Domain	Total accuracy(%)	Hard Undergraduate accuracy (%)	Hard Graduate accuracy (%)	Post-graduate accuracy (%)	Avg. Response Time (s)
GPT-4o	Biology	80	70	85	100	
	Chemistry	33.3	35	35	20	
	Physics	80	90	75	60	
	Total	64.4	65	65	60	
OpenAI-o3	Biology	77.8	80	80	60	36
	Chemistry	55.6	70	40	60	106
	Physics	93.3	95	90	100	33
	Total	75.6	81.7	70	73.3	58
OpenAI-o3 pro	Biology	80	85	80	60	536
	Chemistry	53.3	65	40	60	912
	Physics	91.1	95	90	80	836
	Total	74.8	81.7	70	66.7	795
DeepSeek V3	Biology	75.6	80	80	40	
	Chemistry	44.4	50	45	20	
	Physics	82.2	90	80	60	
	Total	67.4	73.3	68.3	40	
DeepSeek R1	Biology	75.6	75	80	60	134
	Chemistry	60	75	50	40	328
	Physics	84.4	95	85	40	279
	Total	73.3	81.7	70	46.7	247
Human Experts	Biology	62.2	72.5	50	70	1339
	Chemistry	63.3	57.5	65	80	1620
	Physics	62.2	77.5	52.5	40	1936
	Total	62.6	69.2	55.8	63.3	1632

4.1. Overall performance comparison

As illustrated in Figure 1, all evaluated models (GPT-4o, OpenAI-O3, OpenAI-O3 Pro, DeepSeek V3, and DeepSeek R1) achieved higher overall accuracy compared to human experts. This result indicates that contemporary large language models have reached and even slightly surpassed human expert-level performance on logical reasoning tasks. Additionally, internal comparisons within each model family demonstrate that models employing deep-thinking mechanisms generally exhibit superior performance. Specifically, within the GPT series, OpenAI-O3 and OpenAI-O3 Pro clearly outperform GPT-4o, and similarly, DeepSeek R1 exhibits a notable improvement compared to its foundational version, V3. These observations suggest that enhancing logical reasoning mechanisms effectively improves model performance on complex scientific reasoning tasks. However, it is noteworthy that despite the increased processing time designed to facilitate deeper reasoning, OpenAI-O3 Pro achieved nearly identical accuracy to OpenAI-O3, showing no clear performance advantage. The underlying reasons for this unexpected finding will be further analyzed and discussed in subsequent sections.

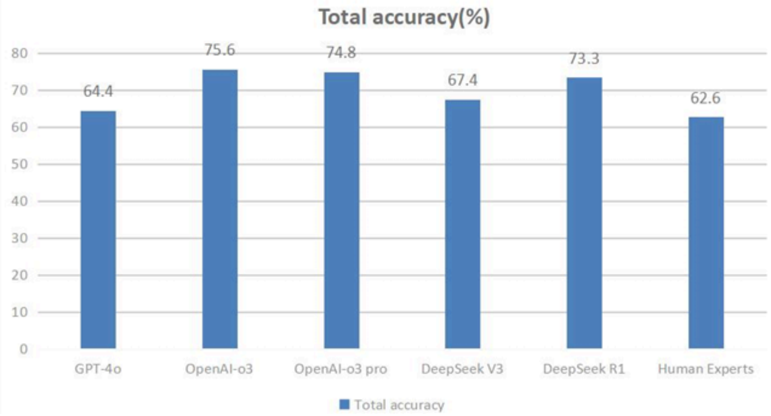

Figure 1. Total accuracy comparison on GPQA

4.2. Response time performance of models

Figure 2 illustrates the average response times of models employing deep-thinking mechanisms (OpenAI-o3, OpenAI-o3 pro, and DeepSeek R1) in comparison to human experts. The results show that the response times of all these models were substantially shorter than those of human experts. Combined with the previously discussed accuracy, these findings further confirm that current large language models outperform human experts in both efficiency and accuracy on logical reasoning tasks.

Specifically, OpenAI-o3 demonstrated the shortest average response time along with the highest accuracy among the evaluated models, indicating an optimal balance between reasoning efficiency and accuracy. DeepSeek R1 exhibited an intermediate average response time—between OpenAI-o3 and OpenAI-o3 pro—while achieving notably higher accuracy compared to its foundational version, DeepSeek V3. This further supports the conclusion that employing deep-thinking mechanisms effectively enhances model accuracy, and emphasizes the importance of appropriately managing reasoning time for optimal efficiency.

However, a notable observation arises concerning OpenAI-o3 pro. Despite significantly extended processing time intended for deeper reasoning, examination of internal logic chains revealed that o3 pro frequently spent additional time retrieving supplementary references or external information related to contextual backgrounds provided in questions. These background details typically served only descriptive purposes and were not directly relevant to the essential knowledge required for problem-solving. The retrieval of this irrelevant information likely contributed to increased response times without a corresponding improvement in accuracy. This phenomenon offers a potential explanation for the nearly identical accuracy observed between OpenAI-o3 pro and OpenAI-o3, an issue that will be further discussed in subsequent sections.

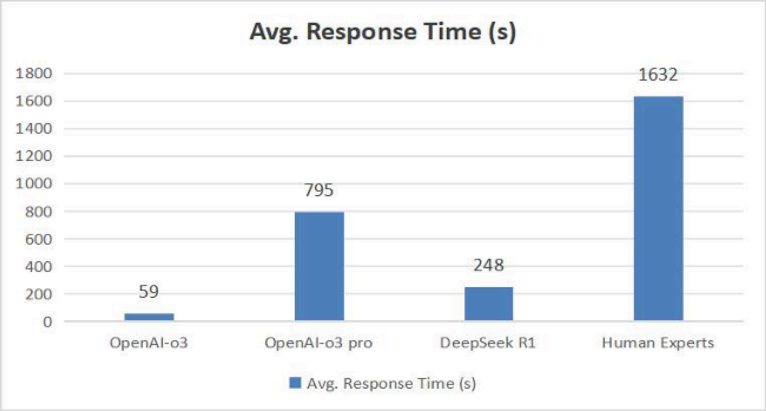

Figure 2. Average response time comparison on GPQA

4.3. Comparison across scientific domains

As shown in Figure 3, when evaluating the performance of models across different scientific domains, it is evident that all models generally achieve higher or comparable accuracy in physics relative to biology, while their performance in chemistry is notably weaker. Overall, the models demonstrate their best performance on physics questions and poorest on chemistry questions. Interestingly, chemistry is also the only domain in which all models underperform relative to human experts. By contrast, the performance of the human expert group remains relatively consistent across different scientific domains, exhibiting minimal variation.

Additionally, the introduction of reasoning mechanisms resulted in inconsistent improvements across different domains. In physics, reasoning significantly enhanced the performance of OpenAI models, whereas DeepSeek models did not exhibit notable improvements. In chemistry, however, the introduction of reasoning mechanisms markedly improved the performance of all models. Conversely, in biology, the reasoning mechanisms did not lead to consistent improvements, and in some cases even resulted in performance degradation.

To further investigate the underlying causes of these performance differences, a detailed analysis of the specific characteristics of the questions in the GPQA dataset was conducted. Physics problems were predominantly structured and computational, typically solvable through direct application of standard formulas, thereby requiring relatively straightforward logical inference [21]. Consequently, the effectiveness of reasoning mechanisms may vary among different models on such comparatively simpler tasks. In contrast, chemistry problems, particularly those involving organic reaction inference, generally demand more complex and deeper logical reasoning, involving multiple iterative reasoning steps and significantly higher computational complexity [22]. As a result, the reasoning mechanisms are likely to be more effective in addressing these complex tasks, explaining their substantial contribution to improved model performance in chemistry. Biology questions, while also involving complex reasoning tasks such as gene sequence inference, tend to be more diverse, and these tasks represent a smaller proportion of the overall domain [23]. Hence, reasoning mechanisms may not consistently yield positive outcomes in biology and could occasionally negatively impact performance on specific questions.

Overall, our analysis indicates that the applicability and effectiveness of reasoning mechanisms within current large language models differ significantly across scientific domains when handling complex logical reasoning tasks. This observation highlights substantial opportunities for further improvement and optimization of models and their reasoning strategies in future research.

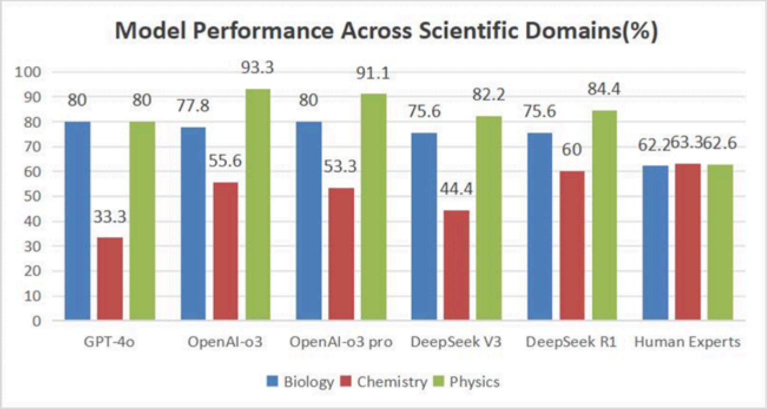

Figure 3. Model performance across scientific domains

Figure 4 further illustrates differences also become apparent when comparing the average response times of deep-thinking models and human experts across these scientific domains. Specifically, among the deep-thinking models, chemistry questions consistently took the longest response times. This observation aligns closely with our earlier analysis of accuracy. Chemistry questions, due to their inherently complex logical reasoning demands, tend to consume more processing time, and current AI models still face significant challenges when handling tasks involving intricate logical reasoning, which likely contributes to their comparatively lower accuracy in chemistry.

Interestingly, the trend differs significantly for human experts, who exhibit the longest average response times on physics questions. A plausible explanation for this divergence is that, although physics questions typically involve fewer complex logical reasoning steps, they often require extensive numerical computations. Unlike AI models, human experts find lengthy calculations more time-consuming and error-prone. AI models, on the other hand, handle such computational tasks with relative ease. This fundamental difference in computational capability between humans and AI models thus leads to contrasting time-investment trends across the evaluated scientific domains.

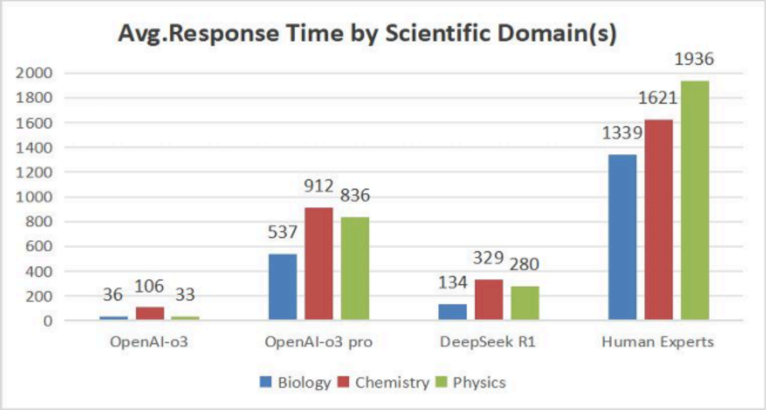

Figure 4. Comparison of average response times across domains

4.4. Performance by question difficulty

As illustrated in Figure 5, clear differences emerge among models when comparing accuracy across varying difficulty levels. The DeepSeek models (DeepSeek V3 and DeepSeek R1) exhibit strong performance on relatively easier questions (Hard Undergraduate level), but their accuracy significantly decreases as question difficulty increases. In contrast, the OpenAI-o3 and OpenAI-o3 pro models, although initially not the top performers at lower difficulty levels, show a distinct improvement in accuracy as difficulty rises. Notably, these two models achieve the highest and second-highest accuracies, respectively, at the most challenging (Post-graduate level or harder) questions. Meanwhile, GPT-4o consistently maintains relatively low accuracy across all difficulty levels.

Overall, these results further confirm that models equipped with deep-thinking mechanisms generally outperform their non-deep-thinking counterparts. Additionally, the GPT series (especially OpenAI-o3 and OpenAI-o3 pro) demonstrates superior performance on highly challenging tasks, whereas the DeepSeek series shows greater stability and reliability on less complex questions.

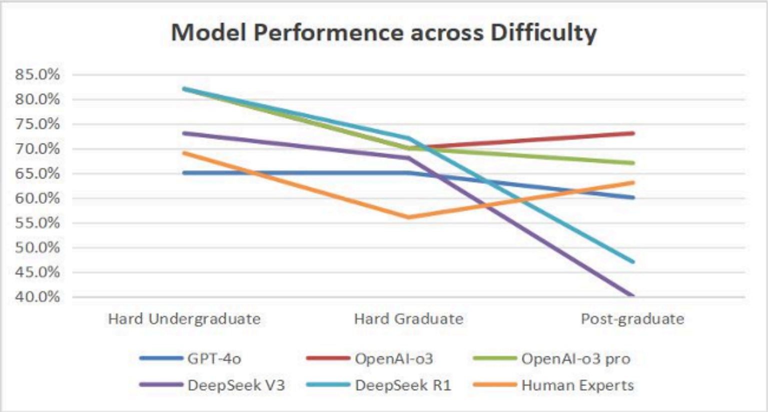

Figure 5. Model accuracy by question difficulty

5. Conclusion

This study systematically evaluated and compared the logical reasoning capabilities of several prominent Large Language Models (LLMs)—GPT-4o, OpenAI-o3, OpenAI-o3 pro, DeepSeek V3, and DeepSeek R1—using the GPQA dataset. Based on our analyses, The following answers are provided to the three core research questions initially posed in the introduction:

Firstly, significant differences in logical reasoning ability among the evaluated LLMs were observed. All tested models collectively outperformed human experts, demonstrating substantial advances in AI capabilities on logical reasoning tasks. Notably, models employing deep-thinking mechanisms (such as OpenAI-o3, OpenAI-o3 pro, and DeepSeek R1) consistently performed better than their counterparts without these mechanisms, highlighting the effectiveness of advanced reasoning strategies.

Secondly, model performance varied distinctly across different scientific domains. Specifically, physics questions, characterized by clearly defined and formula-based computational steps, generally yielded higher accuracy across models. Conversely, accuracy were notably lower in the chemistry domain, which involved more intricate logical reasoning tasks such as organic inference. This observation aligns with prior studies and underscores that current LLMs still face limitations in effectively handling highly structured, multi-step reasoning tasks—an area requiring further improvement.

Thirdly, the activation of deep-thinking functionalities significantly improved model performance on complex reasoning tasks, although the trade-off between increased computational time and improved accuracy warrants further consideration. OpenAI-o3 achieved a more balanced relationship between accuracy and computational cost, whereas OpenAI-o3 pro, despite a substantial increase in processing time, did not significantly outperform OpenAI-o3. This limited improvement may not result from insufficient computational resources but rather from inefficiencies or bottlenecks in the model's reasoning or exploratory pathways, such as unnecessary retrieval of non-essential background information. These findings suggest that future model development should focus not only on enhancing reasoning capabilities but also on optimizing reasoning efficiency.

In summary, this research provides empirical evidence and theoretical insights into both the strengths and limitations of current LLMs on logical reasoning tasks in scientific domains. The findings indicate considerable potential for continued model improvement and highlight areas needing attention, providing valuable insights for future research and development.

References

[1] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. ArXiv, abs/2110.14168. https: //doi.org/10.48550/arXiv.2110.14168

[2] DeepSeek. (2025). DeepSeek R1 Model Update. DeepSeek. https: //api-docs.deepseek.com/zh-cn/news/news250528

[3] Dermata, A., Arhakis, A., Makrygiannakis, M. A., Giannakopoulos, K., & Kaklamanos, E. G. (2025). Evaluating the evidence-based potential of six large language models in paediatric dentistry: a comparative study on generative artificial intelligence. Eur Arch Paediatr Dent, 26(3), 527-535. https: //doi.org/10.1007/s40368-025-01012-x

[4] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. https: //doi.org/10.48550/arXiv.2103.03874

[5] Hsu, C.-Y., Cox, K., Xu, J., Tan, Z., Zhai, T., Hu, M., Pratt, D., Chen, T., Hu, Z., & Ding, Y. (2024). Thought Graph: Generating Thought Process for Biological Reasoning Companion Proceedings of the ACM Web Conference 2024, Singapore, Singapore.https: //doi.org/10.1145/3589335.3651572

[6] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA.https: //doi.org/10.48550/arXiv.2205.11916

[7] Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). Solving quantitative reasoning problems with language models Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA.https: //doi.org/arXiv: 2206.14858

[8] Liu, H., Ding, Y., Fu, Z., Zhang, C., Liu, X., & Zhang, Y. (2025). Evaluating the Logical Reasoning Abilities of Large Reasoning Models. https: //doi.org/10.48550/arXiv.2505.11854

[9] Liu, H., Fu, Z., Ding, M., Ning, R., Zhang, C., Liu, X., & Zhang, Y. (2025). Logical Reasoning in Large Language Models: A Survey. arXiv preprint arXiv: 2502.09100. https: //doi.org/10.48550/arXiv.2502.09100

[10] Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., & Zhang, Y. (2021). LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. https: //doi.org/10.24963/ijcai.2020/501

[11] Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. https: //doi.org/10.24963/ijcai.2020/501

[12] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Comput. Surv., 56(2), Article 30. https: //doi.org/10.1145/3605943

[13] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. ACM Trans. Intell. Syst. Technol. https: //doi.org/10.1145/3744746

[14] OpenAI. (2024). OpenAI O1 System Card. OpenAI. https: //openai.com/o1/

[15] OpenAI. (2025). OpenAI o3 and o4-mini System Card. OpenAI. https: //cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

[16] Ouyang, S., Zhang, Z., Yan, B., Liu, X., Choi, Y., Han, J., & Qin, L. (2024). Structured chemistry reasoning with large language models Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria.https: //doi.org/10.48550/arXiv.2311.09656

[17] Pang, X., Hong, R., Zhou, Z., Lv, F., Yang, X., Liang, Z., Han, B., & Zhang, C. (2025, January). Physics Reasoner: Knowledge-Augmented Reasoning for Solving Physics Problems with Large Language Models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert, Proceedings of the 31st International Conference on Computational Linguistics Abu Dhabi, UAE. https: //doi.org/10.48550/arXiv.2412.13791

[18] Patil, A. (2025). Advancing Reasoning in Large Language Models: Promising Methods and Approaches. https: //doi.org/10.48550/arXiv.2502.03671

[19] Pereira, J., Banchi, L., & Pirandola, S. (2023). Continuous variable port-based teleportation. Journal of Physics A: Mathematical and Theoretical, 57. https: //doi.org/10.1088/1751-8121/ad0ce2

[20] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A., Abid, A., Fisch, A., Brown, A., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A., Safaya, A., Tazarv, A., & Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https: //doi.org/10.48550/arXiv.2206.04615

[21] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA.https: //doi.org/10.48550/arXiv.2201.11903

[22] Wulff, P., Kubsch, M., & Krist, C. (2025). Natural Language Processing and Large Language Models. In P. Wulff, M. Kubsch, & C. Krist (Eds.), Applying Machine Learning in Science Education Research: When, How, and Why? (pp. 117-142). Springer Nature Switzerland. https: //doi.org/10.1007/978-3-031-74227-9_7

[23] Xu, Z., Ding, J., Lou, Y., Zhang, K., Gong, D., & Li, Y. (2025). Socrates or Smartypants: Testing Logic Reasoning Capabilities of Large Language Models with Logic Programming-based Test Oracles. https: //doi.org/10.48550/arXiv.2504.12312