IRUEX: A Study on Large Language Models Problem-Solving Skills in Iran's University Entrance Exam

Hamed Khademi Khaledi, Heshaam Faili
School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran.
{hamed.khaledi, hfaili}@ut.ac.ir

Abstract

In this paper, we present the IRUEX dataset, a novel multiple-choice educational resource specifically designed to evaluate the performance of Large Language Models (LLMs) across seven distinct categories. The dataset contains 868 Iran university entrance exam questions (Konkour) and 36,485 additional questions. Each additional question is accompanied by detailed solutions, and the dataset also includes relevant high school textbooks, providing comprehensive study material. A key feature of IRUEX is its focus on underrepresented languages, particularly assessing problem-solving skills, language proficiency, and reasoning. Our evaluation shows that GPT-4o outperforms the other LLMs tested on the IRUEX dataset. Techniques such as few-shot learning and retrieval-augmented generation (RAG) display varied effects across different categories, highlighting their unique strengths in specific areas. Additionally, a comprehensive user study classifies the errors made by LLMs into ten problem-solving ability categories. The analysis highlights that calculations and linguistic knowledge, particularly in low-resource languages, remain significant weaknesses in current LLMs. IRUEX has the potential to serve as a benchmark for evaluating the reasoning capabilities of LLMs in non-English settings, providing a foundation for improving their performance in diverse languages and contexts.

1 Introduction

Recent advances in Large Language Models (LLMs), particularly exemplified by GPT-4 (Achiam et al., 2023), have dramatically transformed the field of artificial intelligence with profound implications for education. Rapid expansion in computational capacity, model size, and sophistication of underlying algorithms has marked the evolution of these models. With billions of parameters, models like GPT-4 can generate human-like text with an unprecedented level of coherence and contextual understanding (Brown et al., 2020; Bubeck et al., 2023). This remarkable leap forward is mainly due to innovations in transformer architecture (Vaswani et al., 2023) and extensive pre-training on diverse datasets. These have collectively enhanced the models' ability to perform complex tasks, including those in educational contexts (Radford et al., 2019). These advancements have ushered in a new era of AI-driven education, where LLMs facilitate content delivery and enable personalized learning experiences, making education more accessible and adaptable to individual learning needs (Bommasani et al., 2021; Wei et al., 2022a).

Integrating chatbots powered by large language models into educational environments significantly changes how students receive support. These chatbots are used for various academic purposes, from personalized tutoring to administrative assistance and enhancing student engagement (Labadze et al., 2023; Okonkwo and Ade-Ibijola, 2021). One of the most promising uses is personalized tutoring, where chatbots offer tailored aid, allowing students to grasp complex concepts in subjects like mathematics and physics at their own pace (El Janati et al., 2020). In addition, chatbots are increasingly used to manage administrative tasks, such as responding to common queries about course schedules or submission deadlines, thus reducing the workload of educators and administrative staff (Kadir et al., 2023). Moreover, using interactive dialogue systems to engage students enhanced motivation and improved learning outcomes, particularly in blended and online learning environments. (Chen et al., 2023; Hew et al., 2023).

As LLMs become more prevalent in educational settings, rigorous evaluations are essential to assess their effectiveness across various subjects. These evaluations often focus on key metrics such as accuracy, adaptability, and the models' ability to support

Proceedings of the 31st International Conference on Computational Linguistics, pages 6505–6519
January 19–24, 2025.

critical thinking and problem-solving skills (Chollet, 2019). In mathematics, LLMs are evaluated on their capability to understand and generate solutions to complex problems, with studies showing mixed results depending on the complexity of the issues and the model's training (Lample and Charton, 2019; Didolkar et al., 2024). Physics and chemistry assessments often evaluate a model's ability to simulate experiments accurately or explain scientific phenomena clearly (Latif et al., 2024; Bran et al., 2023). In language arts, such as English, LLMs are assessed for their ability to generate coherent and contextually appropriate essays, summarizations, and other literary tasks (Raffel et al., 2020; Han et al., 2024). These evaluations highlight the strengths and limitations of the models, providing crucial information on how LLMs can be best utilized and improved in educational applications (Kasneci et al., 2023; Fagbohun et al., 2024).

In this study, we embarked on a comprehensive journey to assess the problem-solving capabilities of large language models in the context of Iran's university entrance exam (Konkour). A cornerstone of this research was the development of an extensive and diverse dataset meticulously compiled to include 868 questions on seven distinct topics. This dataset encompassed specialized subjects such as mathematics, physics, and chemistry, as well as general subjects such as Persian literature, Arabic and English language, and religious education. Including low-resource languages like Persian and Arabic added another layer of complexity, ensuring that the dataset challenged the models' technical skills and ability to process and generate accurate responses in languages with fewer training resources. By replicating the real-world challenges students face during the entrance exam, this dataset provided a robust foundation for evaluating the models' proficiency in understanding and solving complex problems across various academic fields.

The evaluation process involved a detailed assessment of several leading LLMs using the compiled dataset, including GPT (Achiam et al., 2023), Gemini (Team et al., 2023), and LLaMA (Touvron et al., 2023). Performance metrics revealed insightful data on the accuracy and ranking of each model, allowing us to draw meaningful comparisons between them. Furthermore, a thorough error analysis was conducted to identify common pitfalls and limitations inherent in these models. To push the boundaries of model performance, we also explored advanced techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and few-shot learning (Wang et al., 2020). These methods were tested to enhance the models' accuracy and generalization ability from minimal examples, particularly in low-resource settings. Through this multifaceted approach, the study highlighted LLMs' current strengths and weaknesses in academic problem-solving and paved the way for future innovations in improving their capabilities.

2 Related Work

In recent years, the development and evaluation of large language models have attracted significant attention, especially in dataset design. These datasets are crucial for assessing LLMs' various capabilities, ranging from classic natural language processing (NLP) tasks to more specialized domains such as knowledge and information retrieval.

Classic NLP Tasks: Among the most widely recognized is the General Language Understanding Evaluation (GLUE) benchmark (Wang, 2018), which includes a series of tasks like sentiment analysis, textual entailment, and sentence similarity to assess various aspects of language understanding. GLUE has established itself as a standard for evaluating LLMs, with ongoing research driving models to surpass state-of-the-art performance continually. Additionally, specialized datasets like Stanford Sentiment Treebank (SST) (Socher et al., 2013) for sentiment analysis and TREC (Li and Roth, 2002) for text classification have been vital in fine-tuning LLMs for specific applications.

Building on the foundation of GLUE, the SuperGLUE benchmark (Wang et al., 2019) introduces even more challenging tasks, such as coreference resolution and commonsense reasoning, pushing LLMs to achieve a deeper level of language comprehension. This benchmark has set a new standard for advanced language understanding, guiding the development of models that more closely emulate human linguistic capabilities.

Question Answering: The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) is a cornerstone in the field of question-answering (QA). It sets a high standard for language models by requiring them to extract precise answers from context-rich Wikipedia articles. Alongside SQuAD, other QA datasets, such as TriviaQA (Joshi et al., 2017) have further advanced the reasoning capabilities of LLMs.

Beyond these general-purpose benchmarks, specialized datasets have played a pivotal role in expanding the scope of language models, particularly in educational and scientific domains. The ARC dataset (Clark et al., 2018) focuses on scientific reasoning, presenting questions from standardized science exams that require models to apply background knowledge and logical reasoning to arrive at correct answers. Other specialized datasets such as MathQA (Amini et al., 2019), OpenBookQA (Banerjee et al., 2019), SciBERT (Beltagy et al., 2019), SciBench (Wang et al., 2023) and CQuAE (Gerald et al., 2024) have been instrumental in challenging models with complex problems in specific fields such as mathematics.

Knowledge: LAMA (LAnguage Model Analysis) (Petroni et al., 2019) is a dataset designed to evaluate models' ability to recall factual knowledge stored in their parameters without external context, testing understanding across various domains like history, science, and geography. Building on this, KILT (Knowledge Intensive Language Tasks) (Petroni et al., 2020) serves as a benchmark suite for tasks that require models to retrieve and utilize external knowledge, including open-domain question answering, entity linking, and fact verification. KILT emphasizes the integration of retrieval-augmented generation tasks, reflecting the growing focus on models' ability to combine language understanding with knowledge retrieval. Additionally, datasets like WikiHop (Welbl et al., 2018) and TREX (Elsahar et al., 2018) advance benchmarks by incorporating multi-hop reasoning with knowledge retrieval, challenging models to perform sophisticated reasoning across interconnected facts.

Low-Resource Language Benchmarks: Research on underrepresented languages remains a challenge, but datasets like IndicNLP (Kakwani et al., 2020) and the Masakhane Initiative (Orife et al., 2020) have made strides in addressing this gap. They provide resources for Indian and African languages, respectively, enabling the evaluation and development of models in low-resource contexts. Recent studies, such as Abaskohi et al. (2024), have also explored the effectiveness of models like ChatGPT in handling Persian NLP tasks, emphasizing the importance of culturally and linguistically relevant datasets. These multilingual and low-resource benchmarks highlight the need for inclusive datasets that capture linguistic diversity, paving the way for advancements in LLMs' generalization and application across languages.

3 Datasets

In this project, we leverage several datasets, each contributing uniquely to the thorough evaluation process. These datasets are carefully selected to cover a broad spectrum of academic subjects, reflecting both theoretical knowledge and practical applications. They also include preparatory materials that simulate real-world educational practices, ensuring that the evaluation process is rigorous and relevant. This section provides detailed descriptions of the datasets used, underscoring their significance and relevance in assessing the performance of LLMs. All data referenced in this study are available for download at GitHub.[1]

3.1 IRUEX Dataset

The Iran university entrance exam is one of Iran's most competitive and significant exams, with nearly one million students participating annually. As a critical factor in university admissions, the content of this exam is highly pertinent for evaluating LLMs' reasoning and problem-solving abilities. The IRUEX dataset comprises multiple-choice questions in the math group from exams conducted between 2019 and 2023, organized into subject categories such as mathematics, physics, and chemistry and general subjects like Persian literature, Arabic, English language, and Islamic religious education. In 2023, an update reduced the tested subjects to only mathematics, physics, and chemistry, reflecting a shift in the exam's structure.

The dataset is provided in LaTeX format, allowing LLMs to interpret and process mathematical formulas and scientific notations effectively. Additionally, questions containing images are disregarded to assess the language model's comprehension of textual information solely. The dataset was carefully compiled by crawling, ensuring a comprehensive and accurate collection of exam content. To facilitate evaluation and ensure accessibility for LLMs that do not understand Persian, we also prepared the translated version of the IRUEX dataset into English. This translation allows us to analyze the impact of language conversion on model performance. Detailed information on the number of questions per category is available in Table 1. Detailed examples, key topics, and focus areas for each category can be found in Appendix A.

[1] https://github.com/hamedkhaledi/IRUEX-dataset

	Math	Chemistry	Physics	Arabic	English	Religion	Persian Literature
2023	34	26	25	-	-	-	-
2022	46	24	22	25	25	25	25
2021	47	25	26	25	25	25	25
2020	44	32	19	25	25	25	25
2019	46	28	24	25	25	25	25
Total	217	135	116	100	100	100	100
Supplementary Questions	6042	5923	5674	3883	7366	4809	2761

Table 1: Number of questions per category in the IRUEX dataset

3.2 Supplementary Questions

Beyond the IRUEX dataset, students often engage with supplementary questions to prepare for entrance exams. This dataset includes a carefully curated selection of questions classified by varying degrees of difficulty, along with detailed descriptive answers. These resources are precious for few-shot evaluations of LLMs, as the descriptive answers enable the assessment of the model's ability to generate comprehensive responses based on minimal context. The supplementary questions were gathered by crawling various educational platforms and resources frequently used by students. The number of questions per category in this dataset is also detailed in Table 1.

3.3 High School Textbooks

To further enhance the evaluation process, we include a collection of all high school textbooks, which were transformed from PDF files into text format for easier integration and use. These textbooks are essential for retrieval-augmented generation, where LLMs access a broader knowledge base to improve the accuracy and relevance of their generated responses. This dataset ensures that models are tested on exam-style questions and the foundational knowledge taught in high schools, comprehensively evaluating their capabilities across a broad knowledge spectrum.

4 Experiments

To thoroughly assess the performance of the IRUEX dataset across multiple language models, we conducted an extensive evaluation using a range of models, including Gemini-Pro (Team et al., 2023), LLaMA3.1-8B, LLaMA3-70B, LLaMA3.1-70B, LLaMA3.1-405B (Dubey et al., 2024), GPT-3.5-turbo, GPT-4 (Achiam et al., 2023), GPT-4o-mini and GPT-4o. These models were accessed through their respective APIs. Our evaluation strategy employed several distinct methodologies:

Zero-shot Learning: In the zero-shot evaluation, models were tested without prior exposure to similar examples or specific task-related instructions. To enhance the reasoning capabilities of the models, we employed Chain of Thought (CoT) (Wei et al., 2022b) prompting. This technique encourages models to articulate their thought processes, leading to more comprehensive and detailed responses rather than presenting the final answer. This strategy was pivotal in tasks requiring complex reasoning and sequential decision-making.

Few-shot Learning: For the few-shot evaluation (Brown et al., 2020), we selected three example questions and answers to serve as contextual references from the supplementary question dataset for each gold question. These examples were chosen using a Term Frequency-Inverse Document Frequency (tf-idf) (Ramos et al., 2003) approach, which focused on key phrases within the questions to ensure that the selected examples were highly relevant to the target question. We used the Jaccard score (Niwattanakul et al., 2013), selecting three distinct examples related to the main question, all with scores below the 0.4 threshold. This metric measured the overlap of key terms, ensuring that the examples were sufficiently similar to the main question to provide relevant context yet distinct enough to avoid redundancy and overfitting. This balance was essential for maintaining content and phrasing diversity while offering meaningful insights. Although we initially explored embedding-based retrieval, the tf-idf method proved more effective in matching the examples to the questions. GPT-4o, which had the best performance in the zero-shot evaluation, was primarily used for these few-shot assessments. This approach was particularly advantageous in categories where providing limited but precise contextual information significantly boosted the model's performance.

Translation-based Evaluation: Due to the lack of native Persian language support in specific models,

Model	Math	Chemistry	Physics	Arabic	English	Religion	Persian Literature
Zero-shot							
LLaMA3.1-8B	11.34	16.76	25.62	27.00	78.00	26.00	25.00
LLaMA3-70B	31.06	37.47	51.12	49.00	87.00	39.00	22.00
LLaMA3.1-70B	41.37	47.97	61.46	50.00	87.00	46.00	27.00
LLaMA3.1-405B	41.47	49.07	72.32	61.00	94.00	46.00	42.00
GPT-3.5	25.45	29.02	38.35	30.00	72.00	24.00	31.00
GPT-4	30.91	43.94	51.17	61.00	95.00	48.00	38.00
GPT-4o-mini	53.18	49.30	71.54	52.00	88.00	42.00	33.00
GPT-4o	52.22	62.06	79.76	68.00	**96.00**	66.00	39.00
Few-shot							
GPT-4o	**61.13**	61.93	**82.11**	**69.00**	92.00	**72.00**	**51.00**
Translation-based							
Gemini-Pro	30.38	31.15	33.01	-	81.00	-	-
GPT-4o	45.21	48.54	62.27	-	96.00	-	-
Retrieval-Augmented Generation (RAG)							
GPT-4o	48.57	**64.76**	76.8	64.00	95.00	68.00	43.00

Table 2: Accuracy results (%) on the IRUEX dataset. The highest accuracy for each experiment is highlighted in **bold**, while the second highest is underlined. The average accuracy is weighted according to the number of problems in each course.

such as Gemini, we translated the IRUEX dataset into English using the Google Translate API. The translated dataset was then evaluated with GPT-4o. This approach allowed us to measure the impact of translation on model performance, explicitly focusing on categories that do not rely heavily on native language nuances. However, we excluded categories such as Persian literature, Arabic, and religion from this evaluation due to the inherent loss of meaning and cultural context when translating these topics into English.

Retrieval-Augmented Generation (RAG): To enhance the model's ability to draw on external knowledge, we implemented a retrieval-augmented generation (Lewis et al., 2020) approach. We chunked high school textbooks into paragraphs and used the BM25 (Robertson et al., 2009) algorithm to retrieve each question's most contextually similar paragraphs. Two adjacent paragraphs were also included to ensure the model had sufficient context. This retrieved information was then provided as part of the input prompt, enabling the models to generate more informed and accurate responses based on relevant external knowledge.

Models were instructed to provide their final answer on the last line of their responses, a crucial directive for accurately parsing and evaluating their outputs. We set the temperature parameter to 0 for all evaluations to minimize randomness and ensure reproducible and deterministic responses. We meticulously recorded the accuracy of each model across different categories of the IRUEX dataset, with results comprehensively summarized in Table 2, offering a clear comparison of model performances across various tasks. Detailed prompts used during the evaluations are documented in Appendix B to ensure transparency and reproducibility. Also, additional evaluation results from other LLMs can be found in Appendix C.

5 Results Analysis

We reported results from various language models in diverse academic and linguistic tasks. In this analysis, we will examine several key observations, including model performance, technique impact, and areas for improvement.

5.1 Model Analysis

This section highlights critical points extracted from Table 2 under the zero-shot setting.

Model Size Impact: The performance of language models has consistently improved with increased model sizes. Notably, within the LLaMA series, LLaMA3.1-405B significantly outperforms its predecessors, LLaMA3-70B and LLaMA3.1-70B, across all categories. This suggests that larger models are generally more adept at handling complex tasks, as evident in their higher scores across various categories. Interestingly, newer versions of LLaMA consistently outperform earlier ones

with the same number of parameters, highlighting advancements in model architecture and training techniques. The IRUEX dataset can also differentiate among different LLMs, providing valuable insights into their performance and capabilities.

Model Comparison: Comparison of GPT-4 and GPT-4o-mini provides insights into how specific variants address problem-solving and language-based tasks. GPT-4o-mini excels in problem-solving and mathematical calculations, as indicated by its top scores in mathematics, chemistry, and physics. However, it lags in language-based questions, where GPT-4 demonstrates superior performance. This discrepancy suggests that while GPT-4o-mini is optimized for specific computational tasks, GPT-4 offers a more balanced performance across different question types.

Low-Resource Languages: The performance of these models in low-resource languages is particularly noteworthy. Although English scores are generally high, models show weaker results in languages like Arabic and Persian. Despite this, LLaMA3.1-405B outperforms GPT-4o in Persian literature, indicating that model architecture and training data quality play significant roles in handling diverse languages. Persian questions involving ancient poetry and Arabic-specific linguistic constructs presented unique difficulties for LLMs. These challenges often stemmed from nuanced semantic constructs, intricate morphology, and the culturally embedded context required to fully comprehend the text. Such details could provide actionable insights for improving model training and fine-tuning in low-resource language settings.

Mathematics Performance: These models' mathematical and computational abilities are promising but exhibit room for improvement. GPT-4o and its mini variant show strong performance in mathematical tasks, with GPT-4o achieving the highest score in this domain. However, even the top performers leave space for advancements in accuracy and efficiency, suggesting ongoing potential for enhancing mathematical reasoning and calculation capabilities in AI models.

Ranking Achievement: GPT-4o's performance is highlighted by its impressive rank of approximately 500 out of 140,000 math group students in a zero-shot setting. This ranking emphasizes the model's strong generalization abilities and effectiveness in various tasks without extensive fine-tuning, showcasing its high-level performance relative to a large pool of competitors.

5.2 Techniques Impact

Another crucial aspect of the analysis is the influence of the techniques used on the model's performance. These techniques offer different levels of enhancement depending on the task at hand.

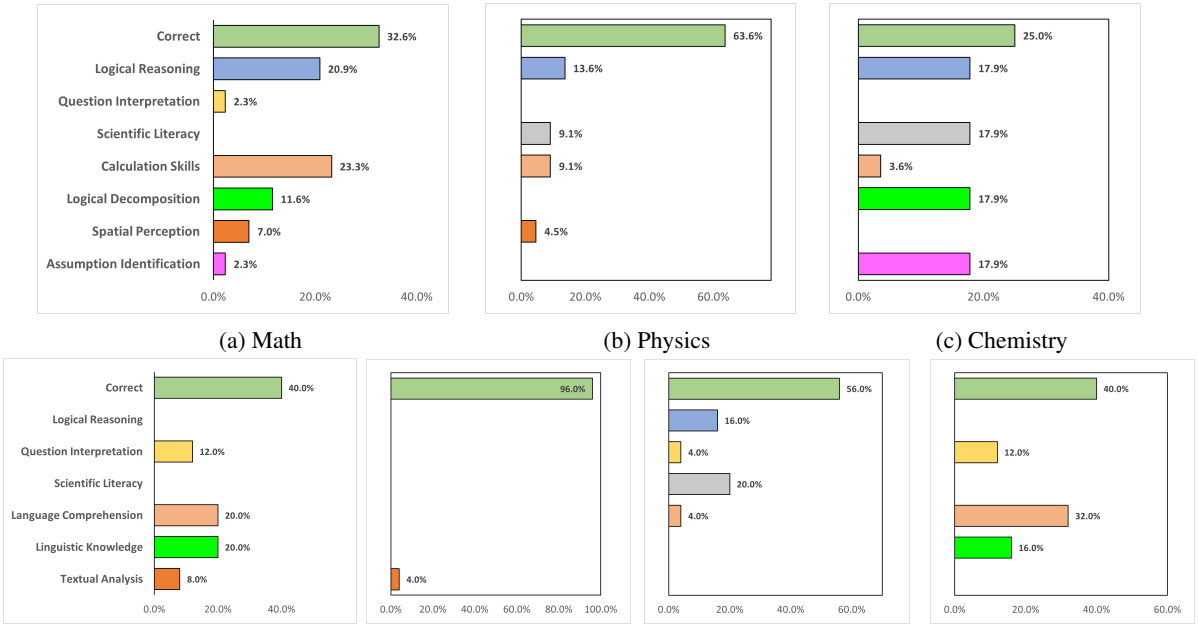
Few-shot learning generally leads to substantial improvements. GPT-4o's performance in a few-shot setting is a testament to this, with notable increases across most subjects, particularly in math (61.13), physics (82.11), religion (72.00), and Persian (51.00). These results suggest that few-shot learning helps the model better understand and respond to complex questions, especially in technical subjects.

RAG also significantly boosts performance, especially in subjects like chemistry, where GPT-4o achieves a top score of 64.76. This technique is particularly effective when models need to retrieve specific information, as seen in the enhanced scores across most categories compared to zero-shot performance. However, RAG's impact on language-dependent tasks is mixed. For example, while it enhances performance in Persian literature and religion, it does not substantially improve results in English or Arabic compared to few-shot learning.

In contrast, translation techniques often degrade performance due to errors introduced during the translation process. The Gemini-Pro model, which relies on translation, consistently underperforms compared to GPT-4o in both technical subjects and English, as indicated by the lower scores across the board. Notably, GPT-4o demonstrates more robust performance in Persian when evaluated directly, without translation. This suggests that translation techniques may not be reliable for tasks requiring precise language understanding, particularly in low-resource languages where translation quality might be inconsistent.

6 Error Analysis

In this section, we analyze the errors made by GPT-4 in a zero-shot setting using the IRUEX dataset. To conduct this analysis, we selected all 192 questions from the 2022 exam to conduct this analysis. Three students who passed the exam with great ranks compared the official solutions to the answers generated by GPT-4. Based on this analysis, we identified the skill deficiencies that contributed to the errors and grouped them into ten categories, with the last six categories adapted from SciBench (Wang et al., 2023). The identified categories are as fol-



(a) Math (b) Physics (c) Chemistry

(d) Arabic (e) English (f) Religion (g) Persian Literature

Figure 1: Error analysis of GPT-4's performance by subject, showing the distribution of correct answers and mistakes across various skill categories.

lows:

- **Question Interpretation**: The ability to fully grasp the question's intent, including understanding key terms and concepts in both the question and its answer choices.
- **Language Comprehension**: The ability to grasp the meaning of individual words or phrases.
- **Linguistic Knowledge**: Proficiency in language-related knowledge, including literary devices, spelling, and grammar.
- **Textual Analysis**: The ability to comprehend and extract relevant information from a given passage of text.
- **Logical Reasoning**: The ability to construct well-reasoned arguments based on observations or assumptions.
- **Logical Decomposition**: The ability to break down a problem into smaller, more manageable parts and understand the relationships between these components.
- **Assumption Identification**: The skill of detecting relevant and necessary assumptions underlying a problem.
- **Calculation Skills**: The capability to accurately perform mathematical operations and computations.
- **Scientific Literacy**: A comprehensive understanding of key scientific principles, terminology, and methodologies across various disciplines.
- **Spatial Perception**: The ability to visualize and understand spatial relationships, which is crucial in fields like physics, chemistry, and geometry.

Based on Figure 1, GPT-4's weaknesses across different areas reveal specific skill deficiencies affecting its performance. In mathematics, a significant challenge for GPT-4 lies in its calculation abilities. The error rate for arithmetic and mathematical computations is notably high, at 23.3%, highlighting the model's difficulty in this domain. Additionally, the construction of coherent arguments and the ability to effectively break down problems into smaller components are other areas of concern, though less prominent. In physics, GPT-4 exhibits fewer calculation errors (9.1%) compared to its math performance, but logical reasoning errors emerge at 13.6%. Spatial perception errors, while not the highest, still affect 4.5% of cases, indicating some difficulty in visualizing physical relationships.

Chemistry presents a different challenge. The model struggles notably with assumption identification (17.9%), logical decomposition (17.9%), and logical reasoning (17.9%). The scientific lit-

eracy category also has a high error rate of 17.9%. However, it is worth noting that RAG improves the model's performance for both chemistry and religion. This suggests that scientific literacy in these domains benefits significantly from enhanced contextual understanding, demonstrating the model's potential for improvement when supplemented with additional information.

The most notable weaknesses in religion are scientific literacy, with a 20.0% rate, and logical reasoning, which accounts for 16.0% of questions. These results suggest difficulties in applying core knowledge and forming well-reasoned arguments in this area. Additionally, question interpretation and language comprehension show smaller but still significant challenges.

In Persian literature, GPT-4 struggles significantly with language comprehension, with an error rate of 32.0%. This indicates a notable deficiency in understanding complex phrases or meanings. Similarly, in Arabic, linguistic knowledge and language comprehension are areas where the model frequently stumbles, with error rates of 20.0% for both. These figures highlight GPT-4's limitations in grasping nuanced meanings in complex linguistic constructs, particularly in non-English languages.

Finally, GPT-4 performs relatively well in English compared to the other domains. Textual analysis shows low error rates (4.0%), indicating that in English language tasks, the model has a firmer grasp of language-related skills. This further emphasizes the difference in performance between English and other languages, particularly when considering more linguistically complex tasks like those found in Persian literature or Arabic.

7 Conclusion

In this paper, we introduce the IRUEX dataset, a novel educational resource designed to assess the capabilities of large language models (LLMs) across seven distinct categories. The dataset is mainly focused on evaluating problem-solving skills and linguistic knowledge, emphasizing underrepresented languages such as Persian and Arabic. By leveraging the IRUEX dataset, we conducted a comprehensive evaluation of various LLMs, employing diverse techniques, and found that the latest, larger models consistently outperformed their predecessors.

Our analysis indicates that while advanced LLMs, like GPT-4o, can achieve performance levels comparable to the top 1% of students, areas remain for improvement, particularly in tasks involving complex calculations and low-resource language processing. We believe the IRUEX benchmark dataset provides a robust foundation for future research aimed at enhancing LLMs' problem-solving and multilingual capabilities. Additionally, we plan to explore the efficacy of different approaches, such as Program-Aided Language (PAL) (Gao et al., 2023), Declarative (He-Yueya et al., 2023) and fine-tuning, on this dataset and across various existing LLMs.

8 Limitations

While the IRUEX dataset provides a valuable tool for evaluating LLMs in problem-solving and linguistic tasks, there are notable limitations that must be considered. First, the dataset is purely text-based, which means it does not evaluate LLMs' multimodal capabilities. As a result, models designed to process textual and visual information cannot fully showcase their strengths in this evaluation. Future iterations of the dataset may benefit from incorporating visual components to provide a more comprehensive assessment of multimodal LLMs.

Additionally, the dataset is restricted to only three languages —English, Persian, and Arabic— limiting its utility in assessing LLMs' multilingual capabilities across a broader spectrum of languages. This focus on a small set of languages, while valuable for exploring underrepresented linguistic contexts, may not capture the full range of challenges faced by LLMs in other low-resource languages. Moreover, the dataset's scope is confined to seven categories of tasks, while comprehensive in some respects, may not cover all possible real-world applications of LLMs, especially in highly specialized or emerging fields. These limitations highlight the need for future dataset versions to include a broader range of languages, modalities, and task categories, enabling a more comprehensive assessment of LLM capabilities.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2189–2203, Torino, Italia. ELRA and ICCL.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. *arXiv preprint arXiv:1907.10738*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Yu Chen, Scott Jensen, Leslie J Albert, Sambhav Gupta, and Terri Lee. 2023. Artificial intelligence (ai) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, 25(1):161–182.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*.

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. *arXiv preprint arXiv:2407.21783*.

Salma El Janati, Abdelilah Maach, and Driss El Ghanami. 2020. Adaptive e-learning ai-powered chatbot based on multimedia indexing. *International Journal of Advanced Computer Science and Applications*, 11(12).

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

O Fagbohun, NP Iduwe, M Abdullahi, A Ifaturoti, and OM Nwanna. 2024. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artifical Intelligence and Machine Learning & Data Science*, 2(1):1–8.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Thomas Gerald, Anne Vilnat, Sofiane Ettayeb, Louis Tamames, and Patrick Paroubek. 2024. Introducing CQuAE : A new French contextualised question-answering corpus for the education domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9234–9244, Torino, Italia. ELRA and ICCL.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Tak Yeon Lee, So-Yeon Ahn, and Alice Oh. 2024. RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13666–13676, Torino, Italia. ELRA and ICCL.

Joy He-Yueya, Gabriel Poesia, Rose E Wang, and Noah D Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *arXiv preprint arXiv:2304.09102*.

Khe Foon Hew, Weijiao Huang, Jiahui Du, and Chengyuan Jia. 2023. Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions. *Journal of Computing in Higher Education*, 35(1):40–68.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Rabiah Abdul Kadir, Mohamad Fairus Zulkifli, Sabrina Binti Tiun, Mohd Modi Lakulu, Shaidah Jusoh, and Ahmad Faridz Ahmad Faudzi. 2023. Educhat: Ai-powered chatbot with personalized engagement for online learning. In *Intelligent Systems Conference*, pages 589–597. Springer.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of ai chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(1):56.

Guillaume Lample and François Charton. 2019. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*.

Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. 2024. Physicsassistant: An llm-powered interactive learning robot for physics lab investigations. *arXiv preprint arXiv:2403.18721*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2:100033.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI blog*, 1(2).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

A Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

A IRUEX description

Sample questions from the IRUEX dataset are presented in Table 3. To facilitate a broader understanding of the dataset content, some of the questions have been translated into English. Below is a summary of each course's key topics and focus areas covered in each course:

- **Math**: Covers key concepts such as discrete mathematics, equations, limits, derivatives, logarithms, and geometry, providing a comprehensive foundation in mathematical principles.
- **Physics**: Explores fundamental topics, including thermodynamics, magnetic fields, electricity, energy, dynamics, and waves, emphasizing the core principles of physical sciences.
- **Chemistry**: Focuses on essential areas like stoichiometry, solutions and solvents, the periodic table, and general chemical knowledge, offering a deep dive into chemical sciences.
- **Persian**: Examines the meaning of verses and words, literary devices, and the works of poets, highlighting key aspects of Persian literary heritage.
- **Arabic**: Covers word meanings, passage comprehension, vowel usage, and translation, focusing on linguistic skills and understanding in the Arabic language.
- **Religion**: Provides an overview of Islamic teachings and the Quran, emphasizing religious knowledge and understanding of Islamic principles.
- **English**: Concentrates on vocabulary development, filling in blanks, and passage comprehension, aimed at enhancing proficiency in the English language.

B Prompts

All the prompts used in the experiments are presented in Table 4. This includes the system prompt, zero-shot prompt, few-shot prompt, and retrieval-augmented generation (RAG) prompt.

C Additional Experiments

In this section, we present a comprehensive overview of the additional experiments conducted to evaluate the performance of various large language models, including Qwen, Mixtral, and Reflection. The complete set of experimental results is presented in Table 5.

Math Correct Answer: D

Consider geometric sequences with a natural ratio greater than one that contains 5 terms.
How many such sequences can be found whose sentences are members of the set {1,2,...,100}?
A) 3
B) 4
C) 6
D) 7

Chemistry Correct Answer: A

Which compound's molecular structure does not have a triple bond?
A) {O_2}
B) CO
C) HCN
D) {N_2}

Physics Correct Answer: B

A ball is released from a height h and falls with a constant acceleration g = 10 frac{m}{{s^2}}. If its average velocity at the end frac{3}{4} of the path is 15 frac{m}{s}, how many meters per second is its average velocity throughout the path?
A) 0.5
B) 7.5
C) 10
D) 5/12

Arabic Correct Answer: C

عيّن ما فيه التأكيد:

A)!هل تَعلم أن العاقل لايظلم لأنّه يري نتيجته قريبا

B)!كأنَ الهواءَ باردٌ فيَجب أن نَلبس مَلابسَ خاصَّة

C)!إنَّ الله لايقذف من الخير إلا في قلوب أوليائه

D)أُنشد الشاعر إنشاذا رائعا في مجلسنا

English	Correct Answer: D
We all know that when two people cooperate with each other, they ..... better Ideas. A) found B) give off C) measure D) come up with	

Religion	Correct Answer: C
According to the verses of the Holy Quran, the acceptance of God's Lordship is determined by which speech of a righteous servant? A)إِنَّا هَدَيْنَاهُ السَبيلَ إِمَّا شَاكِرًا وَ إِمَّا كَفُورًا B)اللَه لَا إِلَه إِلَّا هُوَ لَيَجْمَعَنَكمْ إِلَي يَوْمِ الْقِيَامَةِ C)إِنَ صَلَاتِي وَنُسُكِي وَ مَحْيَايَ وَ مَمَاتِي لِلَهْ D)إِنَ الدَأْرَ الْآخِرَةَ لَهِيَ الْحَيَوَانُ لَوْ كَانُوا يَعْلَمُونَ	

Persian Literature	Correct Answer: A
What is the meaning of each of the following words respectively? درع، آورد، بهرام، سپُردن A)زره، نبرد، سيارهٔ مريخ، طي كردن B)كارزار، ميدان نبرد، سيارهٔ مريخ، پيمودن C)خفقان، توشه و اندوخته، سيارهٔ زحل، طي كردن D)جامهٔ جنگي كه از حلقه هاي آهني سازند، جنگ، سياره زحل، رسيدن	

Table 3: Sample question examples from the IRUEX dataset

System Prompt:
Please present a comprehensive and step-by-step solution for a {CATEGORY} problem.

Zero-Shot Prompt:
The given question is in multiple-choice format with options A, B, C, and D.
After solving, conclude the answer by clearly stating and returning only the correct option in the last line,
enclosed in brackets, and following the 'The correct option is (Option-Letter)' format.
For instance:
Question:
[A sample question]
Solution:
[Provide a detailed step-by-step solution]
The correct option is (A)
Question:
{QUESTION}
Solution:

Few-Shot Prompt:
The given question is in multiple-choice format with options A, B, C, and D.
After solving, conclude the answer by clearly stating and returning only the correct option in the last line, enclosed in parentheses,
and following the 'The correct option is (Option-Letter)' format.
Here are some examples:

Example1:
Question:
{EXAMPLE QUESTION1}
Solution:
{EXAMPLE SOLUTION1}
The correct option is ({CORRECT_OPTION1})

Example2:
Question:
{EXAMPLE QUESTION2}
Solution:
{EXAMPLE SOLUTION2}
The correct option is ({CORRECT_OPTION2})

Example3:
Question:
{EXAMPLE QUESTION3}
Solution:
{EXAMPLE SOLUTION3}
The correct option is ({CORRECT_OPTION3})
Question:
{QUESTION}
Solution:

RAG Prompt
The question is in a multiple-choice format with options A, B, C, and D.
Your task is to solve the question and conclude by clearly stating the correct option.
In the final line, return only the correct option enclosed in parentheses in the following format: 'The correct option is (Option-Letter)'.
For example:
Question:
[A sample question]
Solution:
[Provide a detailed step-by-step solution]
The correct option is (A)
Note: You can include the following context if applicable. Using context is optional.
Context:
{CONTEXT}
Question:
{QUESTION}
Solution:

Table 4: Comparison of system, zero-shot, few-shot, and retrieval-augmented generation (RAG) prompts for solving multiple-choice problems

Model	Math	Chemistry	Physics	Arabic	English	Religion	Persian Literature
Zero-shot							
LLaMA3-8B	9.88	8.85	16.08	23.00	59.00	14.00	17.00
LLaMA3.1-8B	11.34	16.76	25.62	27.00	78.00	26.00	25.00
Gemma2-9B-it	21.28	31.61	47.33	41.00	81.00	32.00	18.00
Gemma2-27B-it	27.07	40.66	53.77	50.00	88.00	40.00	32.00
Mixtral-8x22B	29.99	32.33	43.16	44.00	83.00	44.00	22.00
LLaMA3-70B	31.06	37.47	51.12	49.00	87.00	39.00	22.00
Reflection-LLaMA3.1-70B	36.53	33.32	53.16	38.00	78.00	32.00	30.00
LLaMA3.1-70B	41.37	47.97	61.46	50.00	87.00	46.00	27.00
Qwen2-72B	45.14	45.85	64.13	55.00	93.00	43.00	37.00
LLaMA3.1-405B	41.47	49.07	72.32	61.00	94.00	46.00	42.00
GPT-3.5	25.45	29.02	38.35	30.00	72.00	24.00	31.00
GPT-4	30.91	43.94	51.17	61.00	95.00	48.00	38.00
GPT-4o-mini	<u>53.18</u>	49.30	71.54	52.00	88.00	42.00	33.00
GPT-4o	52.22	<u>62.06</u>	<u>79.76</u>	<u>68.00</u>	**96.00**	66.00	39.00
Few-shot							
GPT-4o	**61.13**	61.93	**82.11**	**69.00**	92.00	**72.00**	**51.00**
Translation-based							
Gemini-Pro	30.38	31.15	33.01	-	81.00	-	-
GPT-4	29.54	35.62	50.77	-	95.00	-	-
GPT-4o	45.21	48.54	62.27	-	96.00	-	-
Retrieval-Augmented Generation (RAG)							
LLaMA3-70B	34.91	39.89	49.74	50.00	86.00	52.00	32.00
GPT-4o	48.57	**64.76**	76.8	64.00	<u>95.00</u>	<u>68.00</u>	<u>43.00</u>

Table 5: Accuracy Results (%) on the IRUEX dataset. The highest accuracy for each experiment is highlighted in **bold**, while the second highest is <u>underlined</u>. The average accuracy is weighted according to the number of problems in each course.