# Evaluating Large Language Models' Proficiency in Answering Arabic GAT Exam Questions

**Mohammad D. Alahmadi**
Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
mdalahmadi@uj.edu.sa

**Mohammed Alharbi**
Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
mnialharbi@uj.edu.sa

**Ahmad Tayeb**
Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
ajtayeb@kau.edu.sa

**Moayad Alshangiti**
Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
mshangiti@uj.edu.sa (corresponding author)

*Received: 24 July 2024 | Revised: 18 August 2024 | Accepted: 22 August 2024*

 *DOI: https://doi.org/10.48084/etasr.8481*

## ABSTRACT

**The Saudi General Aptitude Test (GAT) aims to measure the analytical and inferential learning abilities of high school graduates seeking admission to higher education institutions. Given the need for effective preparation tools, this study investigates the potential of chat generative pre-trained transformers to assist students in preparing for the GAT, especially in Arabic. The primary objective is to assess the effectiveness of Large Language Models (LLMs) in answering questions related to mental and logical abilities, specifically in Arabic. The performance of GPT-4, GPT-4o, and Gemini was examined through 21 experiments to determine their accuracy in answering a range of GAT-related questions. The findings indicate that although GPT-4 and GPT-4o outperformed Gemini in providing accurate answers for the GAT, their current accuracy levels still require improvement.**

***Keywords-ChatGPT; GAT; standardized admissions tests; artificial intelligence; AI-powered tools; machine learning; education; Arabic language***

## I. INTRODUCTION

The Saudi General Aptitude Test (GAT) [1] is a crucial examination for high school graduates in Saudi Arabia who plan to apply to higher education institutions. Administered by the National Center for Assessment (Qiyas), the GAT assesses a variety of learning abilities, including analytical and inferential skills, reasoning, and problem-solving capabilities in both verbal and quantitative domains. Given its significance, students should thoroughly prepare for the GAT to improve their chances of securing a place at their desired universities.

Large Language Models (LLMs), such as Gemini and ChatGPT, have shown considerable potential in educational settings [2, 3]. The performance of these models has been studied on standardized testing exams such as GRE [4, 5], SAT [6, 7], and Medical Licensing Examination [8, 9]. The capacity of LLMs to comprehend and produce text that resembles that of a human makes them valuable tools for personalized learning and academic support [10]. Despite their promising capabilities, uncertainty remains regarding their effectiveness in providing accurate and contextually relevant answers, particularly for the GAT in Arabic. The complexity of GAT questions and the nuances of the Arabic language pose unique challenges that have not been extensively explored in previous studies.

This study carried out an empirical evaluation of three LLMs in answering GAT-related Arabic questions. Their accuracy and quality of responses in seven experimental settings were analyzed to determine their potential as effective preparation tools for students. Through a series of 21 structured experiments, this study sought to provide insight into the strengths and limitations of these LLMs in the context of GAT preparation. The LLMs examined were ChatGPT-4, ChatGPT-4o, and Gemini. The dataset consisted of 400 questions selected and manually transcribed from a well-regarded preparation book, covering various sections, such as Algebra, Geometry, Essays, and Comparisons. The accuracy and quality of the responses from GPT-4, GPT-4o, and Gemini were analyzed using a series of structured experiments to determine their benefits and drawbacks for GAT preparation. For each experiment, specific prompts were formulated, designed to guide the models in selecting the correct answer from the options provided. These prompts varied in their approach, including explicit and implicit contexts, structured response formats, and translation tasks, to thoroughly assess the models' abilities in handling the diverse types of GAT-related questions. The results showed that ChatGPT outperformed Gemini in all experiments for all categories. The questions and results of the 21 experiments are available in [11].

## II. BACKGROUND AND RELATED WORKS

### A. Saudi General Aptitude Test

The GAT [12] measures learning abilities such as analytical and inferential potentials of high school graduates and those who want to enroll in higher education institutions. The objectives of GAT are:

- Measure understanding and reasoning.
- Measure learnability regardless of one's prowess in any particular field.
- Develop self-learning capabilities in line with the new stage of higher education.
- Measure perceiving logical relationships and deductive and inductive abilities.
- Measure the ability to solve problems based on mathematical concepts.

The target group is high school graduates of all tracks and those wishing to enroll in higher education institutions, as well as anyone who hopes to join institutes that require this test. The service-targeted sectors are universities, colleges, and any entity that requires this test. Each track has its test, each of which has various sections of 24 items of verbal and quantitative questions. Quantitative questions aim to examine mental abilities in data analysis, algebra, geometry, and arithmetic, while verbal questions examine logical relations, such as deduction, induction, contextual error, analogy, and logical comprehension.

### B. Similar Studies on ChatGPT

In [13], the aim was to provide qualitative insight into the problem-solving skills of ChatGPT in the context of clinical decision-making and medical education in the Chinese language. According to the findings, the performance of ChatGPT in Chinese was, compared to its performance in English, inadequate for medical learning and clinical decision-making. On the other hand, ChatGPT showed a high level of internal concordance and produced some significant breakthroughs in the Chinese language. The efficacy of ChatGPT on the Medical College Admission Test (MCAT) was tested in [14]. This standardized test consists of 230 questions with multiple-choice answers. In addition to evaluating critical thinking and reasoning skills, MCAT evaluates a wide variety of competencies in the scientific, social, physiological, and behavioral sciences. Based on its encouraging findings, this study predicted two key uses of ChatGPT and subsequent versions in the field of premedical education. Increasing diversity and improving fairness among premedical students could be achieved using ChatGPT in premedical education, which could be an essential and creative step forward.

NLP models are increasingly being utilized for various educational applications beyond language learning. For instance, in [15], automation of mapping course learning outcomes to program learning outcomes was demonstrated using NLP, emphasizing the potential of AI-driven solutions to enhance educational program evaluation. This aligns with the objective of this study of leveraging LLMs to help students prepare for standardized exams, such as the Saudi GAT. In [16], the performance of ChatGPT on UK standardized admission tests was evaluated to gain a deeper understanding of its potential as a cutting-edge teaching and test preparation tool. This evaluation sought to gain a better understanding of ChatGPT's potential. This study involved the Test of Mathematics for University Admission (TMUA), the BioMedical Admissions Test (BMAT), the Thinking Skills Assessment (TSA), and the Law National Aptitude Test (LNAT). Based on the results of this study, ChatGPT has the potential to function as an additional tool for subjects and examination styles that gauge aptitude, problem-solving, and critical thinking skills, as well as comprehension of the literature. However, the constraints of this technology in areas such as mathematics, scientific comprehension, and programs underscore the necessity of ongoing expansion and integration with traditional teaching methods to fully realize its potential.

Recent research has focused on optimizing NLP solutions across different model scales and datasets. In [10], the GPT and LLaMA-2 models were examined across various scales, showcasing the importance of model tuning and task diversity to improve language understanding abilities. This aligns with this investigation as it also focuses on task-specific performance and the challenges of Arabic language processing. Furthermore, in [17], the efficacy of ChatGPT was examined on the United States Medical Licensing Examination (USMLE), which consisted of three assessments. This was carried out without any further training or reinforcement, and ChatGPT was able to pass all three tests with a performance that was close to the passing level. Additionally, the explanations provided by ChatGPT had an increased level of concordance and comprehension, which was a truly outstanding characteristic.

In [5], the problem-solving features of ChatGPT and its uses in standardized test preparation were examined, with an emphasis on the GRE quantitative exam. The results showed that the accuracy of ChatGPT was statistically improved by adding contextual cues and instructions to the original questions. Using the updated prompts, ChatGPT demonstrated 84% accuracy, compared to 69% using the original data. This study outlined potential directions for timely improvements and highlighted the areas where ChatGPT struggled with specific questions. It also explained how alterations can be useful for studying for standardized tests such as the GRE. In [18], AceGPT was introduced, which aimed at localizing LLMs to perform better on Arabic language tasks. This work emphasized the need for models specifically tuned for linguistic and cultural contexts beyond English. Similarly, in [19], Arabic-centric LLMs were explored, with Jais and Jais-chat showing improved performance in handling Arabic-language queries through instruction tuning and specific adjustments for the Arabic language.

Further advances in Arabic LLM evaluation were made in [20], which proposed the AlGhafa Evaluation Benchmark. This benchmark provides a comprehensive method to evaluate Arabic LLMs in multiple tasks, helping in the objective assessment of their capabilities. In a similar vein, in [21] the ArabicaQA dataset was introduced, designed for Arabic question-answering tasks, helping to evaluate LLM performance specifically in question-answering contexts. In [22], the focus was on improving Arabic information retrieval, particularly in question-answering systems, which has direct implications for models such as GPT-4 and Gemini in their potential application to educational settings.

This study extends these efforts by evaluating the performance of GPT-4, GPT-4o, and Gemini on Arabic GAT questions. While existing studies focus on general NLP tasks, this work specifically examines these models in the context of Arabic standardized tests, providing insight into their potential as tools for exam preparation.

## III. PROPOSED METHOD

### A. Dataset and Selected Book

The dataset was created by selecting and manually transcribing 400 questions from Black Box 105 (3rd edition) [23], which includes sample GAT test questions that help learners evaluate their proficiency and identify weak areas. This book is particularly useful for developing both verbal and quantitative skills at the introductory level. This book includes five different sections to cover the GAT, including Algebra, Geometry, Essay, Statistics and Graphs, and Comparisons. The questions from the Statistics and Graphs section were excluded since they contain images (e.g., charts, figures, tables, etc.) that would be difficult to manually transcribe for LLM evaluations. The questions were randomly selected, with nearly one question per page, totaling 400: 95 Algebra questions, 60 Geometry questions, 160 Essay questions, and 85 Comparison questions. Table I shows samples of the selected questions from each category translated into English (for simplicity, option A is the correct answer for these questions). Note that the original questions are in Arabic.

TABLE I. SAMPLES OF THE SELECTED QUESTIONS

| Category | Question | Options | | | |
|---|---|---|---|---|---|
| | | a | b | c | d |
| Algebra | The average of five schools is 170, so, what is their sum? | 850 | 800 | 170 | 1000 |
| Geometry | A square is made up of two congruent rectangles. The area of one rectangle is 18. Find the area of the square. | 36 | 40 | 50 | 55 |
| Essay | Complete the sequence 5, 10, 7, 3, 4, 1, ... | 7 | 9 | 5 | 13 |
| Comparison | If Khaled is older than Saad, Mahmoud is older than Abdullah, and Saad is older than Abdullah, compare Khaled and Abdullah. | The first value is greater | The second value is greater | The two values are equal | Data are insufficient |

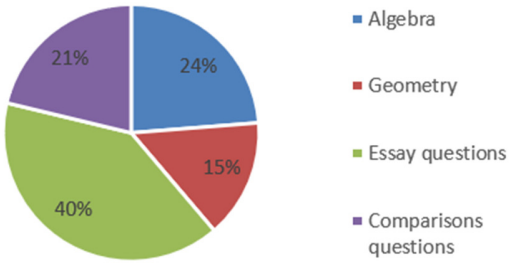


Fig. 1. Distribution of the selected questions.

### B. Experiments

This section describes the experiments conducted to evaluate the performance of the three LLMs (GPT-4, GPT-4o, and Gemini) in answering GAT questions from [23]. The experiments were designed to assess the accuracy and quality of the responses of each model in various scenarios.

### C. Experimental Setup

For each experiment (EXP), the interaction between the language model and the GAT questions was structured using specific prompts. The prompts were designed to guide the model in selecting the correct answer from the options provided. Seven experiments were carried out, which are described in the following.

#### 1) EXP1: Explicit GAT Context

In this experiment, the model was clearly informed that the questions were from the Saudi GAT. The prompt used was:

```
You are about to answer a question from the Saudi
    General Aptitude Test. Please select the correct
    answer from the options provided.

Question: [question]

Answers:
[option1]
[option2]
[option3]
[option4]
```

Fig. 2. Prompt for EXP1.

The motivation for this prompt was to determine whether informing the model that the questions are from a specific test (GAT) affects its accuracy and response quality compared to treating the questions as general knowledge queries.

*2) EXP2: Implicit Context*

In this experiment, the model was not informed that the questions were from the Saudi GAT. The prompt used is shown in Figure 3. The motivation for this prompt was to investigate whether the model's performance differs when it is not informed that the questions are from a specific test (GAT) compared to being explicitly told.

```
Here is a question with four possible answers. Read the
    question and the answers carefully. Then, select
    the correct answer and format your response as
    follows: 'option1: answer', 'option2: answer', etc
    ., where you replace 'optionX' with the correct
    option label and 'answer' with the content of the
    correct option.

Question: [question]

Answers:
[option1]
[option2]
[option3]
[option4]

Select the correct answer and format your response as
    follows: 'option1: answer', 'option2: answer', etc
    ., where you replace 'optionX' with the correct
    option label and 'answer' with the content of the
    correct option.
```

Fig. 3. Prompt for EXP2.

*3) EXP3: Structured Response Format*

Similar to EXP2, but with a different prompt format to ensure the correct answer was identified and formatted correctly. The prompt used is shown in Figure 4. The motivation for this prompt was to assess whether providing specific instructions on how to format the questions and answers affects the model's performance compared to direct questioning without detailed instructions.

```
You will be provided with a question followed by four
    answer options. Carefully read the question and
    each option. Your task is to identify the correct
    answer and format your response using the template
    'optionX: answer text', where 'X' is the number
    corresponding to the correct option, and 'answer
    text' is the exact text of that option.

Question: [question]

Answers:
[option1]
[option2]
[option3]
[option4]

Identify the correct option and format your answer as '
    optionX: answer text'. Ensure 'X' reflects the
    option number, and 'answer text' is copied exactly
    as presented.
```

Fig. 4. Prompt for EXP3.

*4) EXP4: Translation and Answering*

In this experiment, the model was instructed to translate the questions and answers into English before selecting the correct answer. The prompt used is shown in Figure 5. The motivation for this prompt was to determine whether translating questions and answers into English before responding improves the model's accuracy and response quality.

```
You will be given a question and four answer options in
    Arabic. Your task is to first translate the
    question and the answers into English. Then,
    determine the correct answer from the translated
    content. Provide the number of the correct option
    followed by its English translation.

Question: [question]

Answers:
[option1]
[option2]
[option3]
[option4]

Translate the question and answers into English. Then,
    identify the correct option based on the English
    translations. Simply state the option number and
    provide the English translation of the correct
    answer.
```

Fig. 5. Prompt for EXP4.

*5) EXP5: Direct Answering in English*

In this experiment, the model was provided with questions and options in English. The prompt format was the same as EXP3, but the questions were in English. The motivation for this experiment was to evaluate whether providing the model directly with translated questions in English affects its performance compared to previous experiments where translation was part of the process.

*6) EXP6: English Questions with Examples*

The model was provided with a series of questions in English. Examples from different categories were provided to help the model understand the task. The prompt used was:

```
You will be provided with a series of questions, each
    followed by four possible answers. Your task is to
    read the question and options carefully, then
    select the correct answer. Use the template '
    optionX: answer text' for your response, where 'X'
    is the number of the correct option and 'answer
    text' is exactly as it appears.

Below are five examples that illustrate the question
    format and the expected response structure:
[examples]

Please review these examples before proceeding to the
    next question. This will help you understand the
    type of questions and how to format your response
    correctly.

Next question:
[question]

Answers:
[option1]
[option2]
[option3]
[option4]

Identify the correct option and format your answer as '
    optionX: answer text', making sure 'X' correctly
    corresponds to the chosen option, and the 'answer
    text' matches the option precisely.
```

Fig. 6. Prompt for EXP6.

The motivation for this experiment was to assess whether providing the model with question examples in English helps improve its performance in answering subsequent questions.

*7) EXP7: Arabic Questions with Examples*

Similar to EXP6, with the questions and answers provided in Arabic. Examples from different categories (Algebra, Geometry, Essay, and Comparison) were provided to help the model understand the task. The motivation was to investigate whether providing the model with examples of questions in Arabic helps to improve its performance in answering subsequent questions.

### *D. Translation to English*

In addition to the experiments above, GPT-4 was used to translate the questions from Arabic to English for some experiments where the questions and answers are given in English (such as EXP5 and EXP6). The GPT-4 model was instructed to translate the questions and answers into English without selecting an answer. The prompt used was:

```
You will be given a question and four answer options in
     Arabic. Translate the question and each of the
     answer options into English. After translating,
     simply list the translations in the order they were
      provided without selecting an answer.

Question: [question]

Answers:
[option1]
[option2]
[option3]
[option4]

Translate the question and answers into English. List
     the translated question followed by each translated
      answer, numbered accordingly.
```

The OpenAI API was used to interact with GPT-4 to translate the questions and answers from Arabic to English. After obtaining the translations, another author examined their accuracy and found that they were accurate.

### *E. Analysis*

For each experiment, the performance of the models was evaluated based on the accuracy of the answers. The results were compared across different models to identify any significant performance differences. The experiments examined various aspects of the models' capabilities, including their ability to understand and process Arabic questions and provide correct answers in both Arabic and English.

## IV. RESULTS AND DISCUSSION

Figure 7 shows boxplots of the accuracy distributions in the seven experiments and the three large language models used in this study. Detailed results can be found in [11]. The results showed that GPT-4o consistently demonstrated higher accuracy across most experiments compared to the other models. More specifically, GPT-4o achieved the highest median accuracy in EXP 2 (Implicit Context), EXP 3 (Structured Response Format), and EXP 4 (Translation and Answering). GPT-4 also performed well, showing relatively high accuracy in EXP 1 (Explicit GAT Context), EXP 5 (Direct Answering in English), and EXP 7 (Arabic Questions with Examples), although with slightly more variability. The better performance of GPT-4o in 5 of the 7 experiments can likely be attributed to enhancements in its model architecture and training processes. Specifically, GPT-4o may benefit from more advanced attention mechanisms and better management of long-sequence inputs, which are particularly effective in comprehension and context-heavy tasks. Additionally, improvements in fine-tuning techniques and exposure to a more diverse or task-specific training dataset may have contributed to its ability to generalize more effectively across a wider variety of question types. Despite these advances, GPT-4o continues to struggle with geometry-based problems, suggesting that further refinement is necessary in areas that require spatial reasoning and mathematical understanding. Future research could explore the integration of specialized models or more targeted fine-tuning methods to further enhance the model's performance in these challenging areas.

The Gemini model had lower accuracy. The results suggest that both GPT-4 and GPT-4o can provide valuable support in preparation for the GAT, with GPT-4o showing the most promise overall. The consistent underperformance of Gemini in geometry-based questions appears to be linked to its limitations in spatial reasoning and problem-solving. While language models such as Gemini excel in linguistic tasks, they traditionally struggle with tasks that require an understanding of spatial relationships and visual reasoning, which are core components of geometry problems. Unlike language-based tasks, geometry questions require a model to process spatial information and reason about shapes and their properties, areas where Gemini's architecture may not be optimized. The emphasis on linguistic patterns in its training data likely contributes to this performance gap, as the model is not specifically trained to handle spatial or diagrammatic inputs.

Figure 3 shows the accuracy of the answers in the seven experiments and the four categories for the three LLMs. These results confirm that GPT-4 and GPT-4o consistently demonstrated higher accuracy in most categories and experiments compared to Gemini. Specifically, GPT-4 achieved the highest median accuracy in the Algebra and Essay categories. GPT-4o showed strong performance in the Geometry and Comparison category, while both GPT-4 and GPT-4o performed similarly in Algebra. The Gemini model, although generally less accurate than GPT-4 and GPT-4o, still showed reasonable performance, particularly in the Algebra and Comparison categories. These results suggest that both GPT-4 and GPT-4o can provide valuable support in preparing for the GAT in all categories, outperforming the Gemini model.

Although there is no specific passing score for the GAT, according to the Education and Training Evaluation Commission (ETEC), most students score between 55 and 75, with an average score of around 65 [24]. This suggests that the accuracy levels achieved by these models may not yet be sufficient to reliably help students reach or surpass the average. Consequently, while GPT-4 and GPT-4o demonstrate a potential to help with exam preparation, further improvements are necessary for these models to serve as effective tools for preparation for the GAT.

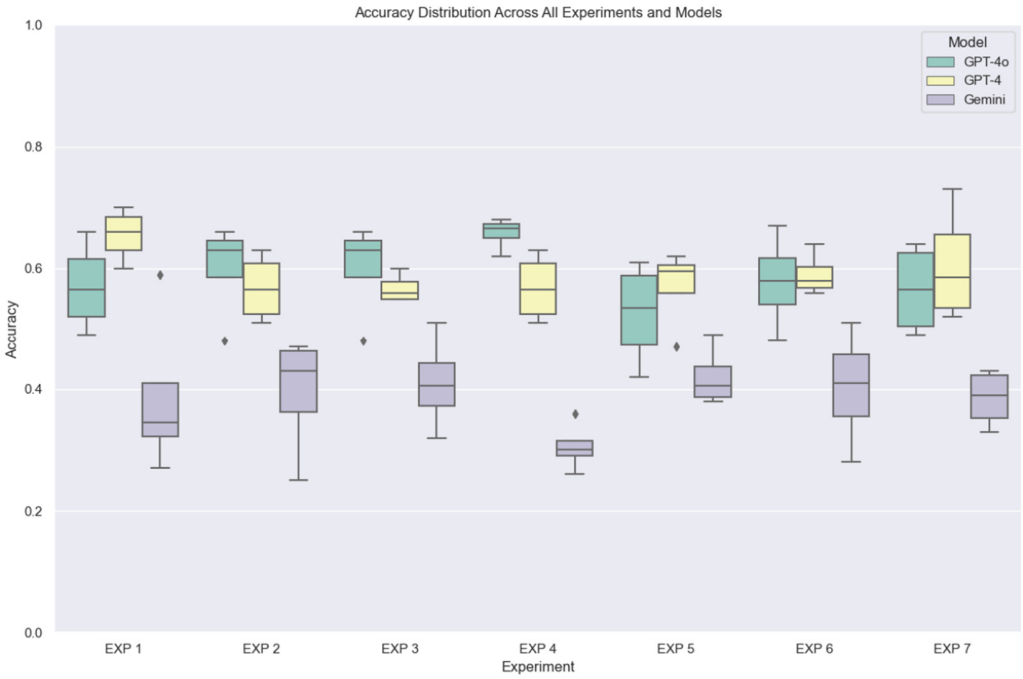


Fig. 7. Accuracy distribution of GPT-4, GPT-4o, and Gemini models across the seven experiments (EXP 1 to EXP 7).

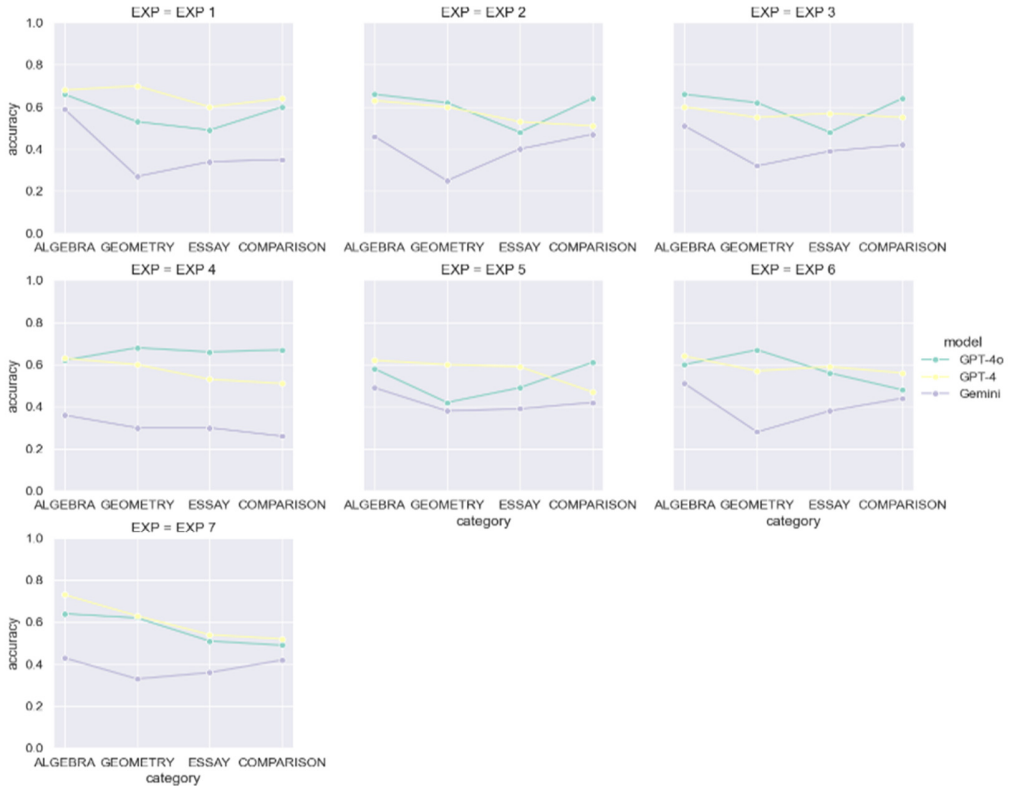


Fig. 8. Accuracy of answers generated by GPT-4, GPT-4o, and Gemini across four GAT question categories: Algebra, Geometry, Essay, and Comparison.

## V. CONCLUSION AND FUTURE WORK

This study evaluated the performance of ChatGPT and Gemini in answering Saudi GAT questions. The accuracy of the responses of GPT-4, GPT-4o, and Gemini was evaluated using a dataset of 400 manually transcribed questions from a test bank book. The structured experiments, which utilized

various prompt formulations, assessed the models' capabilities across different GAT question categories. The novelty of this study lies in its empirical evaluation of LLMs specifically in the context of Arabic-language GAT preparation. Although prior research examined the performance of LLMs in various standardized tests, such as GRE, SAT, and medical licensing exams, this work is unique in focusing on GAT-related questions, which present specific linguistic and cultural challenges. This study highlights the potential of LLMs, especially GPT-4 and GPT-4o, to support students in the Arabic educational context, offering new insights into their application in non-English languages.

However, despite the promising results, several limitations remain. The moderate accuracy rates (40-60%) of these LLMs suggest that although they can support exam preparation, they are not yet reliable as standalone preparation tools. The performance of the models, particularly in geometry and comprehension tasks, reveals gaps in their ability to handle more complex or abstract reasoning questions. Additionally, the reliance on machine translation for some tasks may have affected the accuracy of the models' responses, and further research is needed to assess whether human or alternative LLM-based translation methods could improve performance. Future research should focus on investigating the quality of explanations provided by these models when answering questions, as this can significantly affect students' learning experience and outcomes.

## ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-23-AKSPE-23). Therefore, the authors thank the University of Jeddah for its technical and financial support.

## REFERENCES

[1] "Qiyas General Aplitude Test," *National Center for Assessment*. https://www.etec.gov.sa/en/qiyas.

[2] M. Sullivan, A. Kelly, and P. Mclaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," *Journal of Applied Learning & Teaching*, Jan. 2023, https://doi.org/10.37074/jalt.2023.6.1.17.

[3] K. Malinka, M. Peresíni, A. Firc, O. Hujnák, and F. Janus, "On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?," in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, Turku, Finland, Jun. 2023, pp. 47–53, https://doi.org/10.1145/3587102.3588827.

[4] K. Uludag and M. Zhao, "Can ChatGPT Answer GRE Psychology Questions?" SSRN, Apr. 11, 2023, https://doi.org/10.2139/ssrn.4416365.

[5] U. Farooq and S. Anwar, "ChatGPT Performance on Standardized Testing Exam -- A Proposed Strategy for Learners." arXiv, Sep. 25, 2023, https://doi.org/10.48550/arXiv.2309.14519.

[6] W. Yeadon and D. P. Halliday, "Exploring Durham University Physics exams with Large Language Models." arXiv, Jun. 27, 2023, https://doi.org/10.48550/arXiv.2306.15609.

[7] J. Patel, P. Z. Robinson, E. A. Illing, and B. P. Anthony, "Is ChatGPT smarter than Otolaryngology trainees? A comparison study of board style exam questions." medRxiv, Jun. 18, 2024, https://doi.org/10.1101/2024.06.16.24308998.

[8] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "ChatGPT passing USMLE shines a spotlight on the flaws of medical education," *PLOS Digital Health*, vol. 2, no. 2, 2023, Art. no. e0000205, https://doi.org/10.1371/journal.pdig.0000205.

[9] A. Gilson *et al.*, "How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment." medRxiv, Dec. 26, 2022, https://doi.org/10.1101/2022.12.23.22283901.

[10] A. Kumar, R. Sharma, and P. Bedi, "Towards Optimal NLP Solutions: Analyzing GPT and LLaMA-2 Models Across Model Scale, Dataset Size, and Task Diversity," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14219–14224, Jun. 2024, https://doi.org/10.48084/etasr.7200.

[11] M. Alahmadi, "Evaluating Large Language Models' Proficiency in Answering Arabic GAT Exam Questions." Zenodo, Jul. 24, 2024, https://doi.org/10.5281/ZENODO.12803333.

[12] Qiyas - General Aptitude Test." https://etec.gov.sa/en/service/generalabilitytest/notes.

[13] X. Liu *et al.*, "Performance of ChatGPT on Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese." medRxiv, Apr. 18, 2023, https://doi.org/10.1101/2023.04.12.23288452.

[14] V. L. Bommineni, S. Bhagwagar, D. Balcarcel, C. Davatzikos, and D. Boyer, "Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning." medRxiv, Jun. 06, 2023, https://doi.org/10.1101/2023.03.05.23286533.

[15] N. Zaki, S. Turaev, K. Shuaib, A. Krishnan, and E. Mohamed, "Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation," *Education and Information Technologies*, vol. 28, no. 12, pp. 16723–16742, Dec. 2023, https://doi.org/10.1007/s10639-023-11877-4.

[16] P. Giannos and O. Delardas, "Performance of ChatGPT on UK Standardized Admission Tests: Insights From the BMAT, TMUA, LNAT, and TSA Examinations," *JMIR Medical Education*, vol. 9, no. 1, Apr. 2023, Art. no. e47737, https://doi.org/10.2196/47737.

[17] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, 2023, Art. no. e0000198, https://doi.org/10.1371/journal.pdig.0000198.

[18] H. Huang *et al.*, "AceGPT, Localizing Large Language Models in Arabic." arXiv, Apr. 02, 2024, https://doi.org/10.48550/arXiv.2309.12053.

[19] N. Sengupta *et al.*, "Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models." arXiv, Sep. 29, 2023, https://doi.org/10.48550/arXiv.2308.16149.

[20] E. Almazrouei *et al.*, "AlGhafa Evaluation Benchmark for Arabic Language Models," in *Proceedings of ArabicNLP 2023*, Sep. 2023, pp. 244–275, https://doi.org/10.18653/v1/2023.arabicnlp-1.21.

[21] A. Abdallah *et al.*, "ArabicaQA: A Comprehensive Dataset for Arabic Question Answering," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington, DC, USA, Jul. 2024, pp. 2049–2059, https://doi.org/10.1145/3626772.3657889.

[22] M. Alghamdi, M. Abushawarib, M. Ellouh, M. Ghaleb, and M. Felemban, "Enhancing Arabic Information Retrieval for Question Answering," in *Proceedings of the 7th International Conference on Future Networks and Distributed Systems*, Dubai, United Arab Emirates, Dec. 2023, pp. 366–371, https://doi.org/10.1145/3644713.3644763.

[23] N. I. A. Hafeez, *Black box 105, 3*rd ed. Saudi Arabia: Nabaa Printing and Distribution, 2021.

[24] ETEC, "Open Data," *Eucation and Training Evaluation Commission - ETEC*. https://etec.gov.sa.