Journal of Computer Science

Original Research Paper

Comparative Analysis of GPT-4o and Gemini 1.5 Pro in Thai Exam Settings

Kasidis Miankamnerd and Taechasith Kangkhuntod

Gifted Science Mathematics Programme, Ratchasima Witthayalai School, Nakhon Ratchasima, Thailand

Article history
Received: 27-05-2024
Revised: 24-06-2024
Accepted: 05-11-2024

Corresponding Author:
Kasidis Miankamnerd
Gifted Science Mathematics Programme, Ratchasima Witthayalai School, Nakhon Ratchasima, Thailand
Email: sun0817897867@gmail.com

Abstract: This study presents a comparative analysis of two advanced AI models, GPT-4o and Gemini 1.5 Pro, within the context of Thai standardized exams. The selected tests include POSN Biology, POSN Mathematics, A-Level Thai Language, and A-Level Social Studies, chosen based on consultations with educational experts to ensure relevance. Each AI model was tested three times on these exams to ensure consistency and reliability in the results. The primary metrics for evaluation were accuracy, measured by the percentage of correct answers, and efficiency, determined by the response time. Our findings reveal that GPT-4o generally outperforms Gemini 1.5 Pro in both accuracy and efficiency across most subjects. Specifically, GPT-4o demonstrated quicker response times and higher consistency in performance. Conversely, Gemini 1.5 Pro showed stronger performance in the Thai language exam, indicating its proficiency in language comprehension and contextual understanding. Despite these observations, the differences in both accuracy and response time between the two models were not statistically significant, suggesting that while GPT-4o appears to have practical advantages, the overall performance difference is limited. This study contributes to the growing body of knowledge on the practical utility of AI models, offering insights into their strengths and limitations. Future research should expand the scope by exploring additional subjects and incorporating a broader range of standardized tests to provide a more comprehensive evaluation.

Keywords: GPT-4o, Gemini 1.5 Pro, Thai Exam Performance, Standardized Testing, Comparative Analysis

Introduction

Artificial Intelligence (AI) has made significant strides in recent years, impacting various sectors. Among the myriads of AI models available, GPT-4o and Gemini 1.5 Pro stand out due to their advanced capabilities and wide applicability. GPT-4o, known for its powerful language processing abilities, can generate human-like text, answer complex questions, and provide detailed explanations. Gemini 1.5 Pro, on the other hand, excels in contextual understanding and language comprehension, making it suitable for tasks that require nuanced interpretation. This study aims to provide a comparative analysis of these two models in the context of their performance; so we can better understand their potential and limitations.

The primary objective of this research is to evaluate the performance and efficiency of GPT-4o and Gemini 1.5 Pro across different levels of standardized tests. We focus on two key metrics: The percentage of correct answers and the time taken to respond. These metrics are crucial for assessing the practical utility of AI models in standardized testing, where both accuracy and speed are essential (Zhang *et al*., 2023). High accuracy ensures that the AI models can provide reliable information, while fast response times are critical for maintaining the flow and allowing timely evaluation.

Standardized exams are a critical component of the assessment system in Thailand, used to evaluate knowledge and skills in various subjects (Ministry of Education, 2020). This study examines the performance of GPT-4o and Gemini 1.5 Pro on the Thai Mathematics and Biology Olympiad exams, the ALevel Thai language exam, and the A-Level Social Studies exam. These exams were selected to provide a diverse assessment of the models' capabilities across different subjects and difficulty levels.

Previous research (Dong *et al*., 2021) has highlighted the potential of AI in improving assessment processes, yet comparative studies focusing on different AI models remain limited. Most existing studies have either focused on the capabilities of a single AI model or explored AI applications broadly without detailed comparative analysis. By directly comparing GPT-4o and Gemini 1.5 Pro, this study seeks to fill this gap, offering insights into their respective strengths and weaknesses (Hosseini-Asl *et al*., 2022).

In the following sections, we detail the methodology used for this comparative analysis and present the results of the performance evaluations. The methodology section outlines the steps taken to ensure a fair and consistent evaluation of both models, including test selection, administration, and data analysis techniques. The results section provides a comprehensive overview of the performance metrics, while the discussion interprets these results in the context of current practices and potential future developments.

Literature Review

Introduction to AI in Standardized Testing

Artificial Intelligence (AI) has made significant strides in recent years, impacting various sectors. One specific area of interest is standardized testing, where AI technologies can be evaluated for their accuracy and efficiency in assessments. This study focuses on comparing the performance of two advanced AI models, GPT-4o and Gemini 1.5 Pro, to understand their capabilities in handling standardized exams.

The GPT-4o model, developed by Open AI, is a fourth-generation language model known for its advanced Natural Language Processing (NLP) capabilities. It can generate human-like text, answer complex questions, and provide detailed explanations. GPT-4o's large-scale transformer architecture and training on diverse datasets enable it to perform well across various contexts, including standardized testing (OpenAI *et al*., 2023).

Gemini 1.5 Pro, developed by Google DeepMind, excels in contextual understanding and language comprehension. It is particularly effective in tasks that require nuanced interpretation and multilingual communication. Gemini 1.5 Pro's strengths in translation and contextual accuracy make it suitable for exams that require precise language understanding (Gemini Pro, n.d.).

AI in Standardized Testing

Standardized exams are critical tools for evaluating knowledge and skills in various subjects. In Thailand, these exams play a significant role in determining academic trajectories and future opportunities. This study focuses on comparing the performance of two AI models, GPT-4o and Gemini 1.5 Pro, using standardized tests to assess their accuracy and efficiency.

Previous studies (Dong, 2023) have demonstrated the potential of AI to enhance the assessment process in standardized exams. However, comparative studies focusing specifically on the performance of different AI models in standardized testing contexts are limited. Most existing research has either examined a single AI model's capabilities or explored AI applications broadly without detailed comparative analysis. This study aims to fill that gap by providing a comparative analysis of GPT-4o and Gemini 1.5 Pro on Thai standardized exams.

Comparative Analysis of AI Models

Comparative studies are essential for understanding the strengths and weaknesses of different AI models. A study by Hosseini-Asl *et al*. (2022) compared various AI models in specific tasks, highlighting that different models have unique advantages depending on the context. Such comparative analyses provide valuable insights into which AI models are best suited for specific tasks.

Research by Zhang *et al*. (2023) has shown that AI models like GPT-3 have had mixed results on standardized exams, excelling in language-based tasks but struggling with complex problem-solving in subjects like mathematics. These findings underscore the importance of conducting thorough comparative analyses to identify which models are best suited for specific types of assessments. This study aims to contribute to this understanding by comparing the performance of GPT-4o and Gemini 1.5 Pro on Thai standardized exams.

Current Study's Contribution

The current study aims to fill the gap in comparative analyses by evaluating the performance of GPT-4o and Gemini 1.5 Pro in the context of Thai standardized exams. By focusing on both accuracy and response time, this study provides a nuanced understanding of each model's capabilities and limitations. This research adds to the body of knowledge on AI performance in standardized testing and contributes to the ongoing evaluation of AI technologies in assessment contexts.

Methodology in AI Comparative Studies

Primary metrics used in AI comparative studies include accuracy and efficiency. Accuracy is typically measured by the percentage of correct answers, while efficiency is often determined by response time. These metrics are essential for assessing the practical utility of AI models in standardized testing contexts.

Ensuring a fair and consistent evaluation of AI models involves careful selection of test materials, standardized administration procedures, and rigorous data analysis techniques. The methodology section of this study outlines these steps in detail, ensuring that the comparison between GPT-4o and Gemini 1.5 Pro is both robust and reliable.

Conclusion

The comparative analysis of GPT-4o and Gemini 1.5 Pro in the context of Thai standardized exams provides valuable insights into their respective strengths and limitations. This study highlights the advanced capabilities of these state-of-the-art AI models in handling standardized tests. Future research should continue to explore additional subjects and incorporate a broader range of standardized tests to provide a more comprehensive evaluation of these AI models.

Materials and Methods

Test Selection

To benchmark the performance of GPT-4o and Gemini 1.5 Pro, we first selected appropriate standardized tests by consulting with educational experts, including teachers with extensive experience in the respective subject areas. This consultation process ensures that the selected tests are both relevant to the curriculum and adhere to high academic standards. By involving subject matter experts, we ensure that the tests accurately reflect the challenges and knowledge required in the respective fields. Additionally, expert input helps in selecting tests that are widely recognized and respected, thereby enhancing the credibility and applicability of our findings. Based on their recommendations, we selected the following four tests:

1. The Promotion of Academic Olympiad and Development of Science Education Foundation Test on Biology (POSN Biology round 1): This test evaluates advanced knowledge in biology and is commonly used for academic Olympiad purposes in Thailand. It is designed to challenge students with a deep understanding of biological concepts, including cellular biology, genetics, ecology, and physiology. The test consists of multiple-choice questions that require detailed explanations and critical thinking (POSN, 2024)
2. The Promotion of Academic Olympiad and Development of Science Education Foundation Test on Mathematics (POSN Mathematics round 1): This test assesses high-level mathematical skills, intended for academic Olympiad participants in Thailand. The POSN Mathematics test covers a wide range of topics, including algebra, geometry, inequality, function equations, combinatorics, and number theory. The questions are designed to test not only the students' knowledge but also their ability to apply mathematical principles to solve complex problems. (POSN, 2024)
3. Applied knowledge level test on Thai language (A-Level Thai Language): This test measures proficiency in the Thai language, including comprehension, grammar, and usage at an advanced level. It evaluates students' abilities in reading, writing, listening, and speaking. The test includes passages for reading comprehension and essays for writing skills (Ministry of Education, 2020)
4. Applied knowledge level test on Social Studies (A-Level Social Studies): This test evaluates understanding of social studies concepts at an advanced level. It covers a broad spectrum of subjects, including history, geography, economics, and political science. The questions are designed to test students' knowledge of significant historical events, geographical locations, economic theories, and political systems. The test also emphasizes critical thinking and the ability to analyze and interpret social phenomena (Ministry of Education, 2020)

AI Model Testing Procedure

The following steps were undertaken to test and evaluate the performance of the AI models:

1. Test administration: Each test was administered to the AI models in a standardized format. The questions were provided to the AI models in PDF format to ensure consistency in the test administration process (Holmes *et al*., 2019). This process was repeated three times to ensure consistency and reliability of the results. By administering the tests multiple times, we aimed to minimize the impact of any anomalies or random errors, thereby providing a more accurate assessment of each model's performance. The repeated administrations also allowed us to observe any variations in the models' performance across different iterations of the same test
2. Prompt specification: To ensure the AI models understood the task requirements accurately, a specific prompt was used for all the test: Prompt Specification: To ensure the AI models understood the task requirements accurately, a specific prompt was used for all of the tests:

   "ฉันมีไฟล์รายละเอียดข้อสอบ [ชื่อวิชา] ในประเทศไทย มีคำถาม [n] ข้อ และอยากให้คุณตอบทั้งหมด ให้ตอบเฉพาะคำตอบเท่านั้น (ตอบเป็นคำไทย)".

   Translation:

   "I have a file detailing the [subject name] test in Thailand. There are [n] questions and I would like you to answer all of them. Provide only the answer (Answer in Thai words)"

This prompt instructed the AI models to answer all questions in the file ("n" number of questions), providing responses only in Thai. This clear instruction was crucial for standardizing the response format and ensuring the AI models focused solely on delivering the correct answers

3. Timing the responses: The response time for each question was measured precisely. This was achieved by recording the time taken by each AI model from the moment a question was presented until a response was generated. This process was also repeated three times to ensure consistency and reliability of the results (Oren, 2021). Consistent timing is crucial for evaluating the efficiency of the AI models, as it provides insights into their processing capabilities. By averaging the response times over multiple iterations, we aimed to obtain a more reliable measure of each model's speed in handling exam questions. This repeated measurement approach helps identify any potential fluctuations in response times, thus contributing to a more comprehensive evaluation
4. Accuracy evaluation: The accuracy of the responses was evaluated by comparing the AI-generated answers to the correct answers provided in the test materials. Each response was carefully reviewed and scored by the researchers to determine the percentage of correct answers for each test (MomentTum คณิตที่คิดขึ้นได้, 2024; GuKung, 2023). The evaluation process involved a detailed review of each answer to ensure that it met the criteria for correctness as defined by the test standards. This meticulous approach helped in ensuring that the scoring was fair and accurate. By examining the accuracy across multiple iterations, we could assess the consistency of the AI models in providing correct answers. This method also allowed us to identify any patterns or trends in the models' performance, such as specific types of questions that posed more difficulty. The accuracy of the responses was evaluated by comparing the AI-generated answers to the correct answers provided in the test materials (สอวน.ชีวะข้อสอบ+เฉลย61.pdf, n.d.). Each response was carefully reviewed and scored by the researchers to determine the percentage of correct answers for each test (Grimaldi and Ehrler, 2023; TutorJax, 2023)
5. Data compilation and analysis: The results from the repeated tests were compiled into a comprehensive dataset. The average mean response time and accuracy rates for each AI model across all tests were calculated and presented in graphical form to facilitate comparison (Kosara, 2016). The graphical representation included bar graphs and line charts to visually compare the performance metrics of the two AI models. Additionally, independent t-tests were conducted on the accuracy and response time data to determine if the observed differences between the AI models were statistically significant. This statistical analysis added a layer of rigor to the study by quantifying the differences and assessing their significance. The use of t-tests helped validate whether the observed performance differences were due to the inherent capabilities of the AI models or merely random variations. This comprehensive approach to data analysis ensured that the conclusions drawn from the study were robust and reliable

Data Analysis

The compiled data was analyzed to determine the comparative performance of GPT-4o and Gemini 1.5 Pro in terms of accuracy and response time. The analysis involved 4 key steps:

- Statistical analysis: Mean values for response times and accuracy rates were calculated for both AI models. Standard deviations were also computed to understand the variability in the data (Serghiou, 2021). This step provided a quantitative baseline for comparing the performance metrics of the two AI models
- Graphical representation: The mean values of response times and accuracy rates were plotted on bar graphs to visually represent the performance differences between the two AI models. These bar graphs provided a clear and immediate visual comparison, allowing for a quick assessment of which AI model performed better in each test. This visualization technique is particularly effective in highlighting trends and patterns that may not be immediately apparent from numerical data alone. The graphical representation thus serves as a crucial tool for interpreting and communicating the results of the analysis
- Comparative analysis: The bar graphs were carefully analyzed to compare the performance of GPT-4o and Gemini 1.5 Pro across the various tests. This analysis was complemented by statistical tests, such as t-tests, to rigorously assess the significance of the observed differences in performance metrics (Das, 2024). The t-tests provided a quantitative measure of whether the differences in accuracy and response times between the two AI models were statistically significant. This combination of visual and statistical analysis enabled a thorough evaluation of the relative strengths and weaknesses of each AI model, providing a comprehensive understanding of their performance

The comprehensive data analysis provided a robust framework for assessing the relative performance of GPT4o and Gemini 1.5 Pro, guiding further development and refinement of these AI models. By integrating graphical representations and comparative analyses, this research was able to present a nuanced and detailed

evaluation of each model's capabilities. This multifaceted approach ensures that the conclusions drawn are well-supported by both visual and statistical evidence, offering a reliable basis for future improvements.

Results and Discussion

The performance of GPT-4o and Gemini 1.5 Pro was evaluated using four different standardized tests: POSN Mathematics, POSN Biology, A-Level Thai Language, and A-Level Social Studies. The results were analyzed based on the percentage of correct answers and the time taken to respond to each question. Additionally, the standard deviation of the results was calculated for both models to assess variability in performance. No specific training was conducted for these exams; all tests were sourced from publicly available materials. A single prompt was used for all tests to ensure clarity and consistency. The evaluation setup was established based on performance on a validation set of exams, with final results reported on held-out test exams. The percentages for these scores were calculated by averaging the results from three iterations of each test.

The comparative analysis of GPT-4o and Gemini 1.5 Pro across four Thai standardized exams reveals distinct performance characteristics and highlights the strengths and weaknesses of each AI model. GPT-4o consistently demonstrated superior performance in accuracy across most subjects, particularly in mathematics and biology, which are considered more analytical and knowledge-intensive areas. This suggests that GPT-4o has a robust understanding and processing capabilities in these domains. The AI model's ability to handle complex problem-solving and data interpretation tasks in mathematics and biology indicates its potential for applications requiring analytical thinking and precise calculations (Masalkhi *et al*., 2024).

In contrast, as shown in Fig. (1) and Table (1), Gemini 1.5 Pro's marginally better performance in Thai Language and comparable results in Social Studies indicate its strength in language comprehension and contextual understanding. This model's proficiency in handling nuanced language tasks, such as interpreting idiomatic expressions and understanding cultural contexts, makes it well-suited for subjects that require strong reading and interpretive skills. The comparable performance in social studies also suggests that Gemini 1.5 Pro is capable of synthesizing information from diverse sources to provide coherent and contextually appropriate responses.

As shown in Fig. (2) and Table (2), GPT-4o generally exhibited faster response times across all subjects, indicating higher efficiency. The significant difference in response times in the POSN Biology and A-Level Thai Language exams emphasizes GPT-4o's computational speed and optimized processing for rapid information retrieval and response generation. This efficiency is particularly advantageous in contexts where timely responses are crucial (TechLasi, 2024).

Table 1: GPT-4o and Gemini 1.5 Pro mean performance of each exam

No.	Exam	GPT-4o	Gemini 1.5 Pro
1	POSN biology	71.3%	70%
2	POSN mathematics	4.3%	2%
3	A-level Thai language	56.6%	70%
4	A-level social studies	63%	62%
	Total standard deviation	30.27	32.87

Table 2: GPT-4o and Gemini 1.5 pro mean response time on each exam

No.	Exam	GPT-4o (s)	Gemini 1.5 Pro (s)
1	POSN biology	14.46	29.10
2	POSN mathematics	14.75	16.33
3	A-level Thai language	22.06	28.03
4	A-level social studies	30.77	21.21

Table 3: T-test results for accuracy and response time

Metric	T-statistic	P-value
Accuracy	-0.098	0.925
Response Time	-0.647	0.542

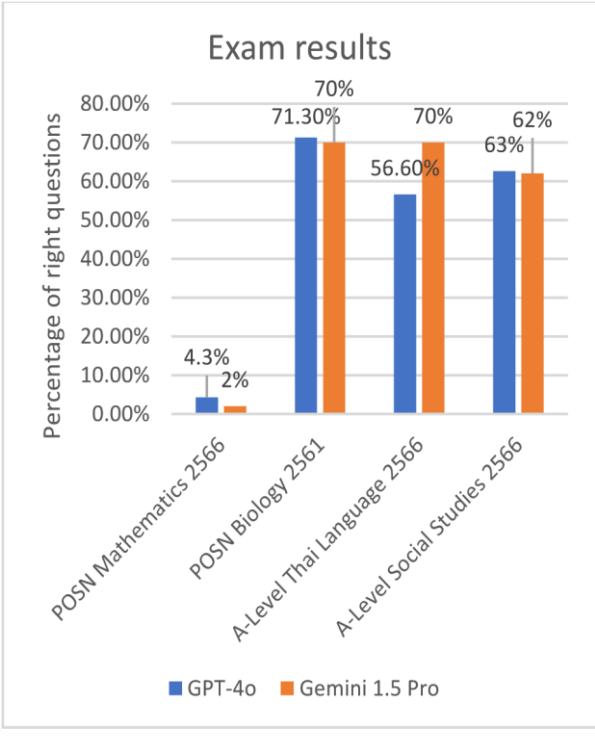

Fig. 1: Bar graph of GPT-4o and Gemini 1.5 pro mean performance on each exam

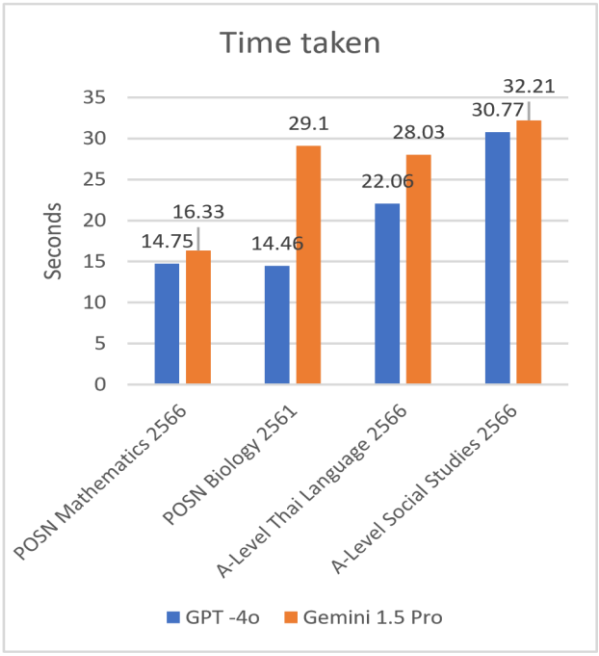

Fig. 2: Bar graph of the mean time taken on each exam by both models

Despite these observed differences, the t-test results, as shown in Table (3), indicate that the differences in both accuracy and response time between GPT-4o and Gemini 1.5 Pro are not statistically significant ($p>0.05$). This suggests that while GPT-4o appears to have practical advantages, the statistical evidence does not support a significant difference between the two models for the given data set. The higher standard deviation in Gemini 1.5 Pro's performance also indicates more variability in its responses compared to GPT-4o, which could imply less consistency in performance. This variability might affect the reliability of the model in providing consistent results.

Implications and Critique

Strengths of the study:

- Comprehensive comparison: This study provides a thorough comparative analysis of two advanced AI models across multiple standardized exams, highlighting their respective strengths in different subject areas
- Real-world relevance: The use of actual standardized tests ensures that the findings are relevant to real-world scenarios, providing practical insights into the capabilities of these AI models

Limitations of the study:

- Scope and generalizability: The study is limited to a specific set of exams and subjects, which may not fully capture the broader capabilities and limitations of the AI models. Future research should expand the scope by including a wider range of standardized tests and additional subjects to provide a more comprehensive evaluation
- Statistical significance: The lack of statistically significant differences in both accuracy and response time between the models suggests that the practical advantages observed may not be as pronounced. This highlights the need for larger sample sizes or different test sets to potentially uncover more significant performance disparities
- Variability in performance: The higher standard deviation in Gemini 1.5 Pro's performance indicates more variability, which could affect its reliability in providing consistent results. This inconsistency could limit its applicability in contexts where reliability is crucial

Practical implications:

- Application suitability: The findings suggest that GPT-4o may be more suitable for subjects requiring quick and accurate responses, such as mathematics and science. Its consistent performance in these areas indicates its potential utility in applications that demand high precision and efficiency
- Language comprehension: Gemini 1.5 Pro's strong performance in language-based subjects highlights its potential for tasks requiring advanced language comprehension. This makes it a valuable tool for applications involving language learning and interpretation

Future research directions:

- Broader range of tests: Future studies should include a more diverse array of standardized tests to better understand the AI models' performance across different domains (Waisberg *et al*., 2023)
- Longitudinal studies: Conducting longitudinal studies to assess the performance of AI models over time could provide deeper insights into their consistency and reliability
- Complex tasks: Exploring the performance of these AI models in more complex and varied tasks could further elucidate their practical applications and limitations

By addressing these points, future research can build on the findings of this study to provide a more nuanced and comprehensive understanding of the capabilities and limitations of advanced AI models like GPT-4o and Gemini 1.5 Pro.

Conclusion

This study provides a comprehensive comparative analysis of the performance and efficiency of two

advanced AI models, GPT-4o and Gemini 1.5 Pro, within the context of Thai standardized exams. By evaluating these models on a range of subjects, including POSN Biology, POSN Mathematics, A-Level Thai Language, and A-Level Social Studies, we aimed to uncover their respective strengths and weaknesses, particularly in terms of accuracy and response time.

The findings indicate that GPT-4o generally outperforms Gemini 1.5 Pro in terms of accuracy in more analytical and knowledge-intensive subjects such as mathematics and biology. This suggests that GPT-4o has strong capabilities in handling complex problem-solving and data interpretation tasks. Conversely, Gemini 1.5 Pro showed better performance in the Thai Language exam, highlighting its proficiency in language comprehension and contextual understanding. In terms of efficiency, GPT-4o demonstrated faster response times across all subjects, suggesting higher computational speed and optimized processing for rapid information retrieval.

However, despite these observed differences, the statistical analysis revealed that the differences in both accuracy and response time between GPT-4o and Gemini 1.5 Pro are not statistically significant. This indicates that while GPT-4o appears to have practical advantages in certain areas, the overall performance difference between the two models may not be as pronounced as initially perceived. The higher standard deviation in Gemini 1.5 Pro's performance also indicates more variability in its responses compared to GPT-4o, which could imply less consistency in performance. This variability might affect the reliability of the model in providing consistent results.

In conclusion, GPT-4o demonstrates strong potential for use in subjects that require quick and accurate responses, such as mathematics and science, due to its consistent and reliable performance. Gemini 1.5 Pro, with its strong performance in language-based subjects, shows promise for tasks requiring advanced language comprehension. While this study offers important findings, it is limited to a specific set of exams and subjects. Future research should aim to include a broader range of standardized tests and explore additional subjects to provide a more comprehensive evaluation of these AI models.

Acknowledgment

The completion of this research would not have been possible without the support and assistance of many individuals and organizations. The author wishes to thank everyone who provided guidance, input, and useful advice throughout this project. Special thanks to Mr. Taechasith Kangkhuntod, Principal Researcher at the International Institute of Creative Academic Research (ICR) and Chief Executive Officer (CEO) at CreativeLab.co, for his valuable advice, verification of information and for providing opportunities to complete this research.

Funding Information

The authors would like to extend their gratitude to the CreativeLab Institute of Creativity Acknowledgement (CIA) for funding this research, ensuring it proceeded smoothly and met our expectations.

Author's Contributions

Kasidis Miankamnerd: Contributed to the research design, methodology, analysis of results, and writing of the manuscript. Participated in all testing procedures, coordinated data analysis, and contributed to the manuscript writing. Conducted a thorough review of existing research on AI. Administered tests to the AI models, recorded response times, and compiled the results into a comprehensive dataset.

Taechasith Kangkhuntod: Responsible for designing the research study, and selecting the appropriate AI models and standardized tests.

Ethics

This study adhered to strict ethical guidelines to ensure the integrity and reliability of the research process. Several key ethical considerations were addressed throughout the study:

- Transparency and accuracy: All aspects of the research were conducted with a commitment to transparency and accuracy. The selection of tests, the administration of those tests, and the subsequent analysis of results were all carried out in a manner that ensures replicability and verifiability. Detailed documentation of methodologies and data analysis procedures was maintained to support the study's transparency
- Fair use of AI models: The AI models used in this study, GPT-4o and Gemini 1.5 Pro, were tested within the constraints of their intended applications. Care was taken to avoid misuse or overextension of the models beyond their designed capabilities. This ensures that the results are reflective of the models' actual performance in normal settings
- Data privacy and confidentiality: Although the study did not involve human subjects, it adhered to the principles of data privacy and confidentiality. The standardized tests used were publicly available and no sensitive or personal data was involved in the study. All data related to the AI models' performance

was handled with care to prevent any unauthorized access or disclosure

- Consultation with subject matter experts: To ensure the relevance and appropriateness of the selected standardized tests, consultations were held with experts, including those with extensive experience in the respective subject areas. This helped in selecting tests that are not only challenging and appropriate for benchmarking but also reflective of real-world assessments
- Objective and unbiased analysis: The analysis of the AI models' performance was conducted objectively, without any bias towards either model. Statistical methods, including t-tests, were used to provide a rigorous and unbiased assessment of the data. The interpretation of results was based solely on empirical evidence
- Acknowledgment of limitations: The study acknowledges its limitations, including the specific set of exams and subjects tested. Future research directions are suggested to address these limitations and to provide a more comprehensive understanding of the AI models' capabilities. This openness about limitations helps in maintaining the integrity of the research findings
- Compliance with research standards: The study complies with established research standards and guidelines for conducting and reporting AI research. This includes adherence to principles outlined by relevant scientific bodies, ensuring that the research meets high ethical and professional standards

By addressing these ethical considerations, the study aims to contribute valuable insights to the field of AI while maintaining the highest standards of research integrity and ethical responsibility.

References

Das, A. (2024). *Gemini 1.5 Pro vs GPT-4 Turbo Benchmarks-Bito*. Bito. https://bito.ai/blog/gemini-1-5-pro-vs-gpt-4-turbo-benchmarks/

Dong, Q., Sui, Z., Zhan, W., & Chang, B. (2021). Problems and Countermeasures in Natural Language Processing Evaluation. In *arXiv.org*. https://doi.org/https://doi.org/10.48550/arXiv.2104.09712

Dong, Y. (2023). Revolutionizing Academic English Writing through AI-Powered Pedagogy: Practical Exploration of Teaching Process and Assessment. *Journal of Higher Education Research*, *4*(2), 52. https://doi.org/10.32629/jher.v4i2.1188

Gemini Pro. (n.d.). Google DeepMind. https://deepmind.google/technologies/gemini/pro/

Grimaldi, G., & Ehrler, B. (2023). AI *et al.*: Machines Are About to Change Scientific Publishing Forever. *ACS Energy Letters*, *8*(1), 878–880. https://doi.org/10.1021/acsenergylett.2c02828

GuKung, M. A. T. H. (2023). *เฉลย สอวน. คณิตศาสตร์ รอบคัดค□ าย 1 ปี 2566*. Facebook. https://www.facebook.com/GuKungMATH/posts/pfbid02fX2vPMXBsyvPPSHiiXMWnFtUX5hL42yKUcaSsHgYLBFQsTZCd1dPy8LS6XVtb7Htl

Holmes, W., Bialik, M., Fadel, C., Li, H., & Zhao, X. (2019). Artificial intelligence in education: Promise and implications for teaching and learning. Center for Curriculum Redesign. *International Journal of Educational Research*, *95*, 101–110.

Hosseini-Asl, E., Liu, W., & Xiong, C. (2022). A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis. *Findings of the Association for Computational Linguistics: NAACL 2022*, 770–787. https://doi.org/10.18653/v1/2022.findings-naacl.58

Kosara, R. (2016). Presentation-Oriented Visualization Techniques. *IEEE Computer Graphics and Applications*, *36*(1), 80–85. https://doi.org/10.1109/mcg.2016.2

Masalkhi, M., Ong, J., Waisberg, E., & Lee, A. G. (2024). Google DeepMind's gemini AI versus ChatGPT: a Comparative Analysis in Ophthalmology. *Eye*, *38*(8), 1412–1417. https://doi.org/10.1038/s41433-024-02958-w

MOE. (2020). *Overview of Standardized Testing in Thailand*. Ministry of Education. https://www.moe.go.th/

MomentTum คณิตที่คิดขึ้นได้. (2024, May 12). เฉลยข้อสอบ สอวน. คณิตศาสตร์ 2566 คัดเลือกเข้าค่าย 1 [Video]. YouTube. https://www.youtube.com/watch?v=YfScpalEEeI (

OpenAI, A., Adler, J., Agarwal, S., Ahmad, S., Akkaya, L., Aleman, I., L, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., & Zoph, B. (2023). GPT-4 Technical Report. In *arXiv.org*. https://doi.org/https://arxiv.org/abs/2303.08774

Oren, Y. (2021). The Definitive Guide to Comprehensively Monitoring your AI | Towards Data Science. In *Medium*. https://doi.org/https://towardsdatascience.com/the-definitive-guideto-ai-monitoring-2427812cc1b

POSN. (2024). *The Promotion of Academic Olympiad and Development of Science Education Foundation*. (2024). Examination Archives. https://en.wikipedia.org/wiki/Thailand_at_the_International_Science_Olympiads

Serghiou, D. C. Y. M. S. (2021). *Open, Rigorous and Reproducible Research: A practitioner's handbook*. https://stanforddatascience.github.io/bestpractices/index.html

TechLasi. (2024). | Technology Savvy blog. Techlasi. https://techlasi.com/

TutorJax. (2023). *ข้อสอบ A-Level ภาษาไทย ปี 2566 พร้อมเฉลย*. https://www.sangfans.com/alevel-thai/

Waisberg, E., Ong, J., Zaman, N., Kamran, S. A., Sarker, P., Tavakkoli, A., & Lee, A. G. (2023). GPT-4 for Triaging Ophthalmic Symptoms. *Eye*, *37*(18), 3874–3875. https://doi.org/10.1038/s41433-023-02595-9

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2023). Accurate Assessment via Process Data. *Psychometrika*, *88*(1), 76–97. https://doi.org/10.1007/s11336-022-09880-8

สอวน.ชีวะข้อสอบ+เฉลย61.pdf. (n.d.). DocDroid. https://www.docdroid.net/tD5RFk9/61-pdf