2024 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT) | 979-8-3503-8918-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLANT63295.2024.00011

IELTS Writing Tasks Evaluation Using LLMs: A Comparative Study of ChatGPT, Claude, Copilot, and Google Gemini

Areeg Fahad Rasheed
College of Information Engineering
Al-Nahrain University
Baghdad, Iraq
areeg.fahad@coie-nahrain.edu.iq

Shimam Amer Chasib
Department of Information Technology
Ministry of Labour and Social Affairs
Baghdad, Iraq
shamamamir2017@gmail.com

Huda Najm Alabbas
Digital Operator Division
Zain Iraq
Baghdad, Iraq
hudanajmalabbas@gmail.com

M. Zarkoosh
Independent Researcher
Baghdad, Iraq
m94zarkoosh@gmail.com

Safa F. Abbas
Department of Cybersecurity Engineering
College of Information Engineering
Al-Nahrain University, Jadriya, Baghdad, Iraq
safaaf.abbas@gmail.com

***Abstract*—IELTS writing tasks are considered one of the most challenging tests among the other three IELTS sections: speaking, reading, and listening. The IELTS Writing section requires significant improvement effort, and obtaining expert feedback can be costly. In recent years, large language models (LLMs), a type of artificial intelligence, have revolutionized and are used in various tasks such as text summarization, question answering, and many others. This paper explores the effectiveness of well-known LLMs in estimating IELTS Writing band scores, offering a comparative analysis of several models, including ChatGPT, Claude, Copilot, and Google Gemini. The data used for this task were collected from IELTS practice test books, and the results were evaluated using the R-square and the average degree of bias. The findings indicate that ChatGPT-4 and ChatGPT-4O showed the closest results to the actual labels and outperformed the other models.**

***Index Terms*—IELTS Writing Tasks, Large Language Models, ChatGPT, Claude, Copilot, Gemini.**

I. Introduction

IELTS is one of the most challenging exams for testing proficiency in English [1], [2]. It consists of four parts: speaking, listening, reading, and writing. Among these, the writing section is particularly challenging. It includes two tasks: Task 1 requires candidates to describe a figure or data representation in at least 150 words. Task 2 involves writing about 250 words to express an opinion on a given topic [3], [4]. Preparing for the IELTS requires significant effort and can be costly, as it often necessitates assistance in evaluating the writing section, which English institutions typically offer. Expert evaluation is not accessible to all students, especially those who can only afford the exam fee [5].

In the last decade, we have witnessed the revolutionary advancements in artificial intelligence (AI), transforming how tasks are performed with minimal or no human intervention. Artificial intelligence has found applications across various fields, including healthcare [6], security [7], [8], education [9], and many others [10].

Large language models (LLMs) are a specialized area within artificial intelligence that utilizes neural networks to generate text that closely resembles human writing [11]. LLMs have significantly impacted various sectors in recent years, influencing data analysis, content creation, and education [11]. These models have been widely adopted in numerous applications, due to their ability to perform a range of tasks effectively [12], [13]. For example, LLMs are used in text classification tasks like identifying fake news and text generation for applications such as chatbots and automated question-answering systems [14]–[18]. Grammar checking and enhancement are also methods used by LLMs [19].

This study aims to assist individuals who are interested to take the IELTS exam but face financial constraints by examining the feasibility of relying on AI models to assess their progress in English writing. The structure of the remainder of this paper is as follows: Section 2 details the models utilized in this study. Section 3 describes the data involved in the evaluation process. Section 4 outlines the methodology employed for assessing tasks. Section 5 discusses the performance metrics used and provides the results obtained from the LLMs. Finally, Section 6 is the conclusion of the study.

II. Large language models LLMs

This section will briefly overview the LLM models used in this study and the main properties.

- ChatGPT: Different versions of ChatGPT have been released in recent months by OpenAI, each with varying capabilities and limitations. Four versions of ChatGPT are

979-8-3503-8918-0/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLANT63295.2024.00011

used in this study [20], [21]. The first model, ChatGPT-3.5, was trained on a massive dataset and can generate diverse types of text based on user queries. It was trained with 175 billion parameters. The next version, ChatGPT-4, is a paid version with more capabilities, optimizations, and greater accuracy compared to ChatGPT-3.5, being trained on about 1 trillion parameters. The third model is ChatGPT-4O, where 'O' stands for 'Omni,' indicating its ability to handle different input types, including text, images, audio, and video, surpassing the previous models. It also comes with improved memory efficiency and the ability to learn from past user interactions, leading to better conversational performance. The fourth model of ChatGPT used in this study is ChatGPT-4O Mini, a lighter version of ChatGPT-4O. It provides a cost-effective and efficient solution, with the limitation of only accepting text input [22], [23].

- Claude: This is an LLM model produced by Anthropic [24] , utilizing a transformer encoder-decoder architecture [25]. Claude employs Constitutional AI [26], an approach devised by Anthropic to provide helpful and harmless content without extensive human feedback. Claude supports multiple tasks like other LLMs, including text generation, classification, and summarization. The first version of Claude launched in March 2023. This paper uses Claude 3.5 Sonnet, the most powerful model in the Claude series, released on June 20, 2024. Claude 3.5 Sonnet accepts images and text as input and generates text as output [27], [28].
- Microsoft Copilot AI: This generative AI model is built on large language models (LLM). Launched by Microsoft in February 2023, it handles various tasks with high efficiency. Copilot AI uses the Microsoft Prometheus model [29], which seamlessly combines Bing search data with outputs from ChatGPT-4o, ChatGPT-4, and ChatGPT 4 turbo for superior performance [30].
- Gemini AI is a large language model (LLM) developed by Google DeepMind and launched on December 6, 2023. It is based on a transformer architecture, specifically a decoder-only model [25]. Gemini AI can be used for various tasks such as providing summaries, translating languages, answering questions, etc. Multiple versions of Gemini AI exist, including Gemini Ultra, Gemini Pro, and Gemini Flash. This study will use Gemini Flash, a lightweight model optimized for speed and efficiency. It supports text, images, and audio [31].

III. Data Collection

This paper aims to evaluate the most well-known large language models (LLMs) in predicting IELTS writing task band scores. For this purpose, we collected 68 tests from official Cambridge practice books, including questions, answers, and expert evaluations. We used the expert evaluations as ground truth labels while the questions and answers were fed to various LLMs for assessment. Among these tests, 34 were related to Task 1, which typically involves describing a figure or diagram, and the remaining 34 focused on Task 2, where candidates are required to provide a 250-word opinion on a given topic. Figure 1 shows the images of the IELTS practice books used to collect the data in this study. Notably, we focused exclusively on academic tests for our analysis. The Cambridge practice books are not freely available, so we have not provided the questions and answers for the tests. Instead, we have provided a spreadsheet [1] file containing multiple columns, which can also be found in Table 2. The first column lists the book number, with samples collected from books 9, 11, 12, 13, 14, 15, 16, 17, and 19. The second column specifies the task, either Task 1 or Task 2. The third column contains the band score provided by Cambridge experts (true label), while the remaining columns contain band score predictions according to the LLMs.

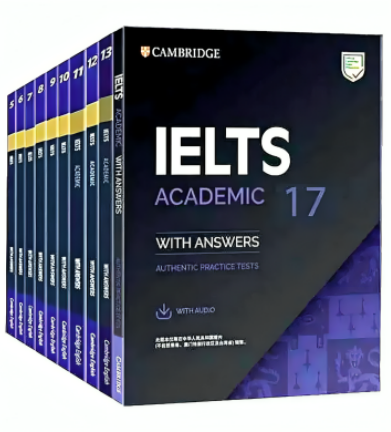

Fig. 1: Cambridge practice tests book

IV. Task Assessment Process

Three components are used to assess each task in the task assessment process. The first component, the prompt, serves as an instruction to the model, informing it of the required task. The second and third components are the tasks' questions and answers (see Figure 2). In this study, we have used two prompts, as illustrated in Table I. The reason for using different prompts is that some models, like Claude and ChatGPT versions 4 and 4o, can accept multiple images. We sent both the question and the answer images, which the model interpreted. In models that do not accept more than one image, such as Gemini and Copilot, we sent only the question, and the answer was fed as text. We provided only the answer text for models that do not accept images.

V. Evaluation metrics and results

We computed the R-squared value and the average degree of bias to quantify the effectiveness of each model in predicting the IELTS band scores. R-squared measures the correlation between the model's predicted band scores and the actual band provided by the Cambridge experts. The average bias measures

[1] https://github.com/AREEG94FAHAD/IELTS_writing_evl

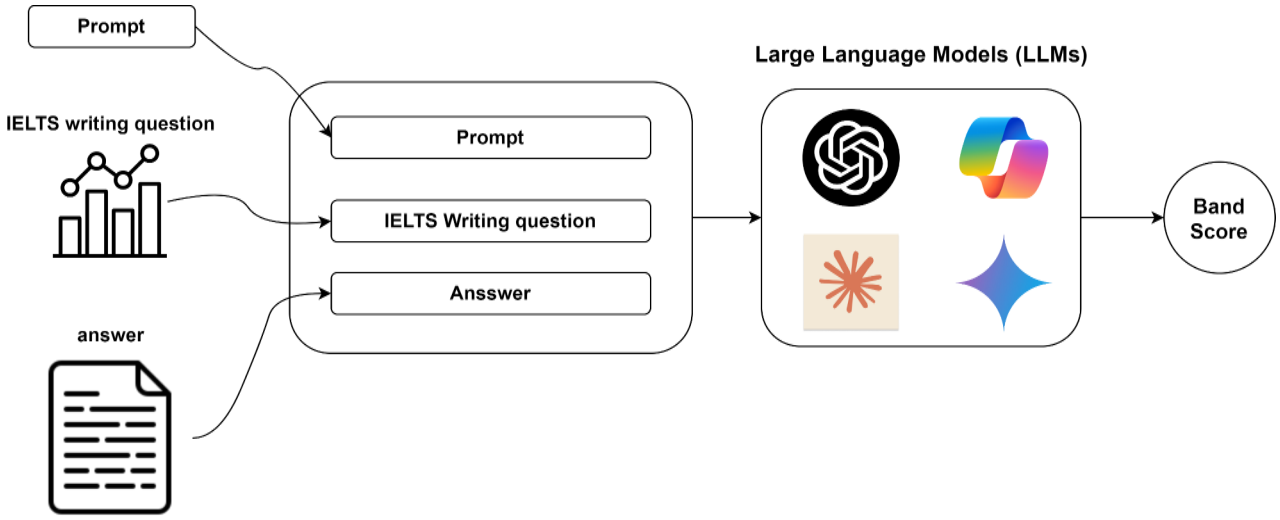

Fig. 2: Evaluation process diagram

TABLE I: Prompts used for different models in the study

Prompt	Models
These two images: one represents the question and the other represents the answer for the IELTS Academic Writing Task. Estimate the band score of the answer out of 9.	Claude, ChatGPT version 4 and 4o
Check this answer for the IELTS writing task and estimate the band score out of 9. {the answer}:	Gemini, Copilot, ChatGPT version 3.5 and 4o-min

the degree of discrepancy between the actual labels and the model's predictions. For example, if the actual label is seven and the model predicts 7.5 or 6.5, this indicates a bias of 0.5. We calculate the average bias across all test samples to assess the model's overall accuracy. We have computed the R-squared and average bias for each task separately.

Figure 3 presents the results of the LLM models evaluating Task 1 and Task 2, respecing the R-Square, respectively. For Task 1, which typically consists of results represented as charts or diagrams, the findings indicate that ChatGPT versions 4 and 4O outperformed the other models by achieving the highest R-squared values. This indices superior image interpretation capabilities in these models. The third-best model in terms of performance was Claude, while the least effective were Copilot and ChatGPT 3.5.

The results for Task 2 show some changes, but ChatGPT-4 and 4O still perform better than the other models. Gemini Flash outperformed Claude, achieving third place in predicting results close to the true labels. Additionally, the Go-Mini and Copilot models performed better in this task than Task 1, as it does not require high image interpretation capabilities. In contrast, the least effective model was ChatGPT 3.5.

Figure 4 represents the average degree of bias concerning Tasks 1 and 2. Generally, the model that outperformed the others is ChatGPT-4, which accurately predicted all the band scores for 34 Task 1 essays. The second-best performer in estimating Task 1 band scores was ChatGPT-4O, with an average bias of 0.221. The third and fourth positions were held by Claude Sonnet and ChatGPT-4O-Mini, with biases of 0.529 and 0.721, respectively. The least accurate models in estimating the band scores were ChatGPT-3.5 and Copilot, which achieved biases of 0.824 and 0.765, respectively.

Considering Task 2, the leading performance, similar to Task 1, was achieved by ChatGPT-4 and ChatGPT-4O, with average biases of 0.088 and 0.294, respectively. The third place was taken by the Gemini Flash model, followed by Claude Sonnet in fourth place. The least effective model for predicting Task 2 was ChatGPT-3.5, with an average bias close to 0.853, which is significantly higher than the other models.

VI. Conclusion

This paper addresses the question: Can LLMs reliably estimate the band scores of proficiency for Writing Tasks 1 and 2? To answer this, we compared well-known LLMs, including ChatGPT (versions 4, 4O, 4O-Mini, 3.5), Claude 3.5 Sonnet, Gemini Flash, and Copilot. The estimated band scores were evaluated against results provided by Cambridge experts, using metrics such as R-squared and average bias degree. The results show that ChatGPT-4 significantly outperformed other models by precisely predicting the outcomes of 34 tests from Task 1, with only a 0.088 bias for Task 2. The second-best model was ChatGPT-4O. These findings suggest that LLMs have the potential to assist individuals in estimating their proficiency in writing, which can help them prepare more effectively before taking the exam and potentially save money.

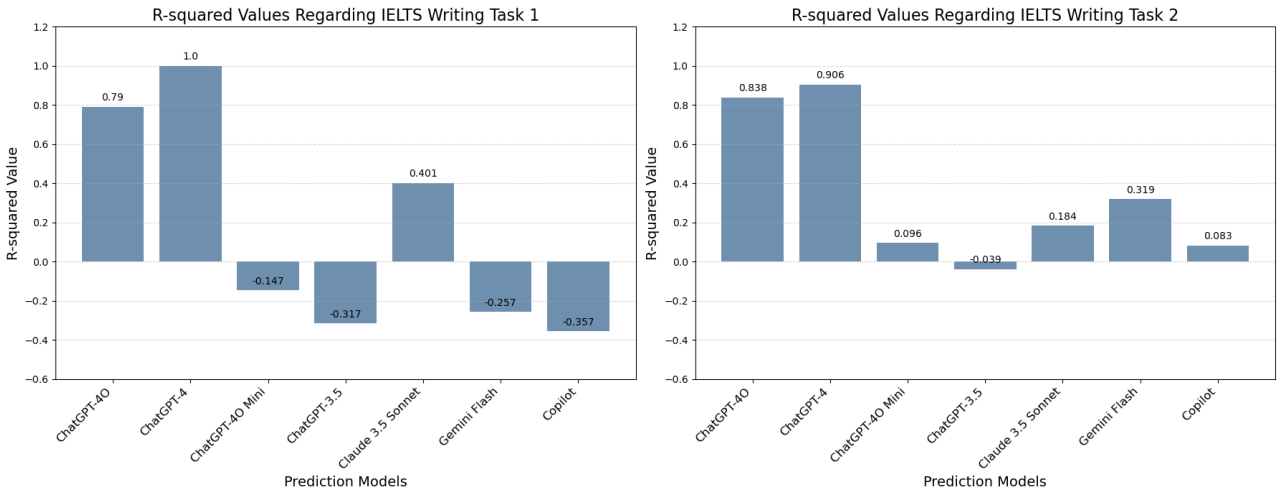

Fig. 3: Models' performance with respect to R-squared regarding IELTS Writing Task 1 and Task 2.

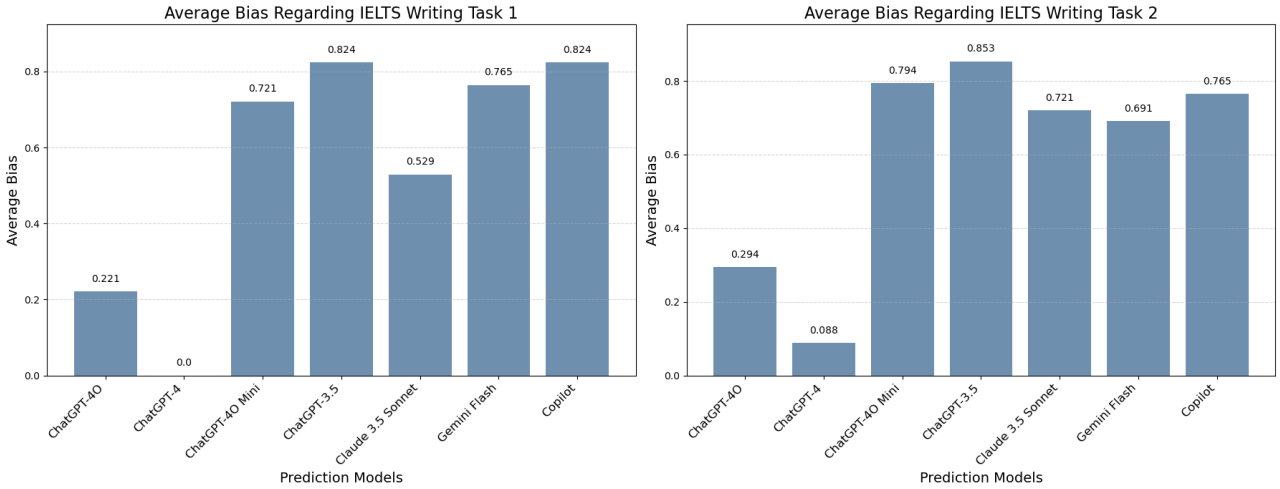

Fig. 4: Average bias for each model regarding IELTS Writing Task 1 and Task 2.

For future works, we suggest trying different prompts, collecting more samples, and investigating the effectiveness of fine-tuning on different machine learning models.

References

[1] N. H. Nguyen and K. D. Nguyen, "Vietnamese learners' performance in the ielts writing task 2," *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task*, vol. 2, pp. 170–189, 2022.

[2] H. H. Uysal, "A critical review of the ielts writing test," *ELT journal*, vol. 64, no. 3, pp. 314–320, 2010.

[3] C. Clapham, *The development of IELTS*. Cambridge University Press, 1996, vol. 4.

[4] A. Hashemi and S. Daneshfar, "A review of the ielts test: Focus on validity, reliability, and washback," *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, vol. 3, no. 1, pp. 39–52, 2018.

[5] W. S. Pearson, "'remark or retake'? a study of candidate performance in ielts and perceptions towards test failure," *Language Testing in Asia*, vol. 9, no. 1, p. 17, 2019.

[6] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[7] A. F. Rasheed, M. Zarkoosh, and F. Elia, "Enhancing graphical password authentication system with deep learning-based arabic digit recognition," *International Journal of Information Technology*, vol. 16, no. 3, pp. 1419–1427, 2024.

[8] A. F. Rasheed, M. Zarkoosh, and S. S. Al-Azzawi, "The impact of feature selection on malware classification using chi-square and machine learning," in *2023 9th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2023, pp. 211–216.

[9] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, and Y. Li, "A review of artificial intelligence (ai) in education from 2010 to 2020," *Complexity*, vol. 2021, no. 1, p. 8812542, 2021.

[10] A. F. Rasheed, M. Zarkoosh, and S. S. Al-Azzawi, "Multi-cnn voting method for improved arabic handwritten digits classification," in *2023 9th International Conference on Computer and Communication Engineering (ICCCE)*, 2023, pp. 205–210.

[11] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and

the ugly," *High-Confidence Computing*, p. 100211, 2024.
[12] B. Meskó, "The impact of multimodal large language models on health care's future," *Journal of medical Internet research*, vol. 25, p. e52865, 2023.
[13] S. Chen, M. Guevara, S. Moningi, F. Hoebers, H. Elhalawani, B. H. Kann, F. E. Chipidza, J. Leeman, H. J. Aerts, T. Miller *et al.*, "The effect of using a large language model to respond to patient messages," *The Lancet Digital Health*, vol. 6, no. 6, pp. e379–e381, 2024.
[14] J. Fields, K. Chovanec, and P. Madiraju, "A survey of text classification with transformers:how wide? how large? how long? how accurate? how expensive? how safe?" *IEEE Access*, 2024.
[15] A. F. Rasheed, M. Zarkoosh, S. F. Abbas, and S. S. Al-Azzawi, "Arabic offensive language classification: Leveraging transformer, lstm, and svm," in *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*. IEEE, 2023, pp. 1–6.
[16] Y. Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li, "Large language model (llm) ai text generation detection based on transformer deep learning algorithm," *arXiv preprint arXiv:2405.06652*, 2024.
[17] A. F. Rasheed and M. Zarkoosh, "Mashee at SemEval-2024 task 8: The impact of samples quality on the performance of in-context learning for machine text classification," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 60–63. [Online]. Available: https://aclanthology.org/2024.semeval-1.10
[18] A. F. Rasheed, M. Zarkoosh, S. F. Abbas, and S. S. Al-Azzawi, "Taskcomplexity: A dataset for task complexity classification with in-context learning, flan-t5 and gpt-4o benchmarks," *arXiv preprint arXiv:2409.20189*, 2024.
[19] L. Netz, J. Reimar, and B. Rumpe, "Using grammar masking to ensure syntactic validity in llm-based modeling tasks," *arXiv preprint arXiv:2407.06146*, 2024.
[20] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of chatgpt: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.
[21] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023.
[22] Future Skills Academy, "Chatgpt versions," 2024, accessed: 2024-08-13. [Online]. Available: https://futureskillsacademy.com/blog/chatgpt-versions/
[23] OpenAI, "Openai," 2024, accessed: 2024-08-13. [Online]. Available: https://openai.com/
[24] Anthropic, "Anthropic," 2024, accessed: 2024-08-13. [Online]. Available: https://www.anthropic.com/
[25] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
[26] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.
[27] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach," *Intelligent Systems with Applications*, vol. 21, p. 200336, 2024.
[28] Claude AI, "Claude ai," 2024, accessed: 2024-08-13. [Online]. Available: https://claude.ai/new
[29] Bing, "Building the new bing," 2023, accessed: 2024-08-13. [Online]. Available: https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing
[30] Microsoft, "Microsoft copilot," 2024, accessed: 2024-08-13. [Online]. Available: https://www.microsoft.com/en-in/microsoft-copilot
[31] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

Table 2: Performance of Different Models on Various Tasks

Task type	Book #	Test #	True label	ChatGPT-4O	ChatGPT-4	ChatGPT-4O Mini	ChatGPT-3.5	Claude 3.5	Gemini Flash	Copilot
1	9	1	7	6.5	7	6.5	6	7	6.5	7
2	9	2	8	7	8	6.5	6.5	7	6.5	7
1	9	3	6	6.5	6	6	6	7.5	6	7
2	9	4	4	5	4	5	4	4.5	5	5
1	11	1	4.5	5	4.5	5.5	3.5	4.5	5	5
2	11	1	5.5	5.5	5.5	5	4	6	6	5
1	11	2	6	6	6	6.5	5.5	6.5	6	5
2	11	2	5	5	5	5	4.5	5.5	6	5
1	11	3	6	6	6	6	5.5	6	6	5
2	11	3	7	7.5	7	6.5	7.5	7	7	8
1	11	4	7	6.5	7	6.5	6	6.5	6	6
2	11	4	5.5	6	5.5	5.5	5.5	6.5	6	6
1	12	1	5	5	5	5.5	4.5	5.5	5	6
2	12	1	6	5.5	6	5	4.5	5	5	6
1	12	2	7	6.5	7	5.5	6.5	6.5	6	6
2	12	2	5	5.5	5	5	5	5	5	6
1	12	3	6.5	7	6.5	6	6	7	6.5	7
2	12	3	7.5	7.5	7.5	7	7	7.5	6.5	8
1	12	4	6	6	6	5	4.5	5.5	5	6
2	12	4	5	5.5	5	6	5	5	5	6
1	13	1	5.5	5.5	5.5	5	5	6	5	7
2	13	1	6.5	6.5	6.5	5	5.5	6	5.5	7
1	13	2	6	5.5	6	4.5	4.5	4.5	4	5
2	13	2	7	7.5	7	5	6.5	6.5	6.5	5
1	13	3	5	5.5	5	4.5	4	4.5	4	7
2	13	3	6	5.5	6	4.5	5	5	5.5	7
1	13	4	6.5	6.5	6.5	6.5	6	6.5	6	7
2	13	4	6	6.5	6	5	4	4.5	6	7
1	14	1	6.5	6.5	6.5	6	6	7.5	6	7
2	14	1	7	6.5	5.5	6	5.5	4.5	6	7
1	14	2	6	6	6	6	5	6.5	6	7
2	14	2	9	8.5	9	8	7	6.5	6.5	7
1	14	3	6	6	6	6	5.5	5.5	6	7
2	14	3	5.5	5.5	5.5	5	4.5	5	5	7
1	14	4	9	8	9	7	7	8	7	7
2	14	4	7	7	7	7	6.5	7.5	6.5	7
1	15	1	6	6	6	6	5.5	5.5	5	6
2	15	1	7	6.5	7	6	5	6.5	6	6
1	15	2	7	6	7	5	5	6	5	6.5
2	15	2	6	6.5	6	6	5.5	7	6.5	6.5
1	15	3	6.5	6.5	6.5	6	6	6.5	6	6.5
2	15	3	7	6.5	7	6.5	6.5	7	7	6.5
1	15	4	6	6	6	5.5	6	6	5	7
2	15	4	6.5	6.5	6.5	5	5.5	5.5	5.5	6.5
1	16	1	5	5	5	4.5	4.5	4.5	4	5
2	16	1	6	6	6	5.5	6.5	7.5	6.5	6.5
1	16	2	6	6	6	5	6	6.5	5	7.5
2	16	2	4.5	5	4.5	4.5	4.5	5	4	7.5
1	16	3	6.5	6.5	6.5	6	5.5	6.5	6	7.5
2	16	3	7	6.5	6	5	5	6	6	6
1	16	4	5.5	5.5	5.5	4	4	5	5	6
2	16	4	4	4	4	3	3.5	3	4	5
1	17	1	6	6	6	5	5.5	6	5	5
2	17	1	6.5	6.5	6.5	6	5.5	6.5	6	7
1	17	2	7.5	7.5	7.5	6.5	6	6.5	6	7
2	17	2	6.5	6.5	6.5	6	5	6.5	6	7
1	17	3	5.5	5.5	5.5	5.5	5.5	6	6	7
2	17	3	6.5	6.5	6.5	6	6.5	7	6.5	6.5
1	17	4	7.5	7.5	7.5	7	7	8	7	6.5
2	17	4	6	6	6	5.5	5	5.5	5	6.5
1	19	1	6.5	6.5	6.5	5	5.5	6	5	6.5
2	19	1	6	6	6	5	5	5.5	5	6.5
1	19	2	6.5	6.5	6.5	6	5	7	6	7
2	19	2	6	6.5	6	6	6	7	5	7
1	19	3	6.5	6	6.5	5	5	5.5	5	7
2	19	3	7	7	6.5	5	6	6	6	7
1	19	4	5	4	5	4	4	4	4	7
2	19	4	6	6	6	5	5	5.5	5	7