ANALYSIS REPORT

Executive Summary:

This report evaluates the performance of three AI language models—Gemini 2.5 Pro, ChatGPT 4o, and MedGemma 4B—in healthcare domain applications. The evaluation used tailored prompts designed with CLEAR, Few-shot, and Chain-of-Thought methodologies to improve statistical reliability. Models were assessed on accuracy, relevance, and domain appropriateness across analysis, generation, and question-answering tasks. ChatGPT 4o emerged as the top performer, followed by Gemini 2.5 Pro, while MedGemma 4B scored lowest, primarily due to its smaller size and local deployment capability. Further research is recommended to assess larger MedGemma models on the same benchmarks to advance insights and inform healthcare AI adoption strategies.

Methodology:

In this phase of the evaluation, three different AI models were tested under controlled conditions to assess their performance on domain-specific prompts. ChatGPT 4o and Gemini 2.5 Pro were run in their respective browser-based environments, while MedGemma 4b was deployed using a Google Colab instance, with executable code provided in the *medgemma.ipynb* file for reproducibility and further analysis. All prompts used for this phase are compiled in the *Phase_2_Answers.md* file, while the complete documentation of results is organized in the *Phase_3_Results.xlsx* file. Outputs generated by the models are made available either as shareable links from the browser instances or as saved code outputs from Colab, ensuring transparency in the evaluation process.

To maintain fairness and eliminate contextual influence, each task prompt was presented to the models in a fresh session, ensuring no prior conversation history could affect the response. The answers were recorded exactly as generated, with no modifications made. Model performance was then scored on a standardized scale: a score of 4 was awarded if the response fully met task requirements with accurate, relevant, and domain-appropriate output; a score of 2 was assigned if the response contained partial correctness, flawed reasoning, or incomplete detail; and a score of 0 was given if the output was incorrect, irrelevant, or logically unsound. This scoring approach allowed for consistent comparison across models. Remarks and justifications for the scoring decisions accompany the results, providing context for observed strengths and weaknesses in each model's performance.

Results:

After employing the evaluation methodology described earlier, the overall performance results for the three AI models—GPT 4o, Gemini 2.5 Pro, and MedGemma 4B—are presented visually in Figure 1. The figure summarizes aggregated scores derived from the dataset in the *Phase_3_Results.xlsx* file, which includes comprehensive assessments across Accuracy, Relevance, and Domain Appropriateness dimensions. These scores were calculated based on each model's performance over nine different prompt combinations spanning three prompt types (CLEAR, Few-Shot, and Chain of Thought) and three task kinds (Classification/Analysis, Generation/Creation, and Question-Answering).

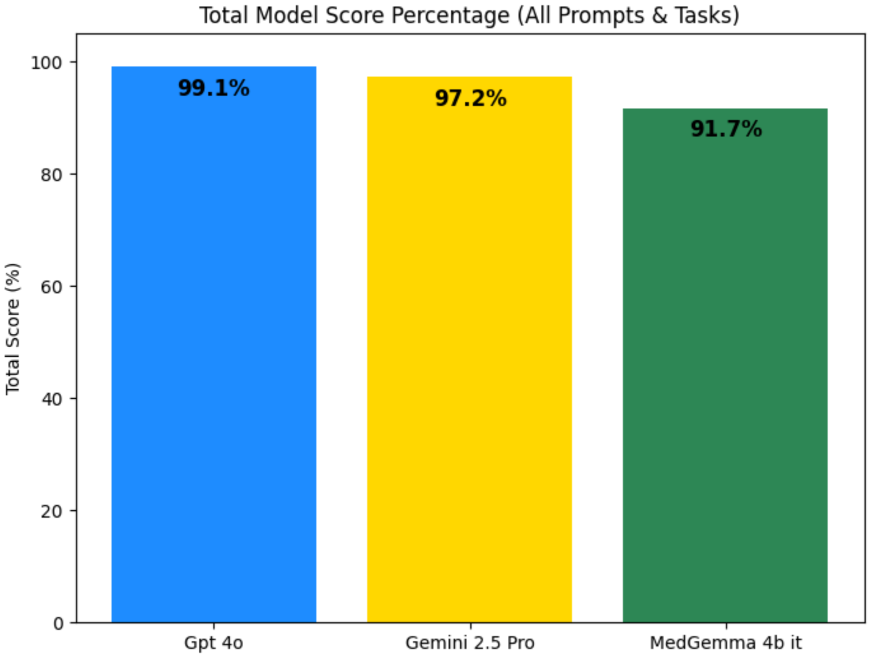

Figure 1 The given shows the overall results of each model based on the scores taken from Phase_3_Results.xlsx

A clear hierarchy in model performance is evident from the results. GPT 4o leads with an impressive total score of 99.1%, closely followed by Gemini 2.5 Pro at 97.2%, while MedGemma 4B records the lowest among the three, scoring 91.7%. This comparative analysis highlights the strengths and limitations of each model given their architectural scale and training resources. Notably, GPT 4o and Gemini 2.5 Pro belong to the class of trillion-parameter models, possessing substantial computational capacity and training data volumes that likely contribute to their superior accuracy, relevance, and appropriateness scores. In contrast, MedGemma 4B, with its relatively compact 4 billion parameters, inherently operates under a more constrained resource and model complexity environment.

Delving into task-specific performances, it becomes clear that MedGemma 4B underperforms most noticeably in the Question-Answering task across prompt types. This area reflects the highest discrepancy when compared to the other two models, suggesting that the reduced number of parameters and architectural design might limit its ability to synthesize contextual information required for high-level domain reasoning. Given that Question-Answering requires integrating diverse knowledge facets and maintaining coherence across responses, MedGemma's limitations highlight the impact of model scale on proficiency in such generative and comprehension-intensive tasks.

Furthermore, the evaluations considered the raw model outputs as generated without any post-processing or bias correction, ensuring unbiased reflection of each model's in-the-moment capabilities with zero context retention between prompts. The methodology enforced the condition that every prompt interaction started with a cleared context to guarantee fairness and isolate prompt-response performance distinctly.

These findings underscore the correlation between model size, parameter count, and resource investment with the effectiveness of producing accurate, relevant, and domain-appropriate outputs. While MedGemma 4B provides strong foundational results and remains competitive, especially considering its smaller architecture, further scaling or refinement might be necessary to match the top-tier performance observed in GPT 4o and Gemini 2.5 Pro.

In conclusion, the aggregated scores and qualitative remarks derived from the *Phase_3_Results.xlsx* file paint a comprehensive picture of the comparative abilities of these modern AI models. The results provide valuable insights for selecting models based on task demands, resource availability, and performance needs, with GPT 4o and Gemini 2.5 proving advantageous for demanding and critical domain applications, while MedGemma 4B offers resource-efficient capabilities suited for contexts with computational constraints.

Discussion:

The results demonstrate clear performance distinctions among the three evaluated models, highlighting critical trade-offs between scale, resource availability, and real-world applicability. GPT 4o's top score reflects its extensive parameter count and broad training, enabling it to deliver highly accurate, relevant, and domain-appropriate responses consistently across all prompt types and tasks. Similarly, Gemini 2.5 Pro performs closely behind GPT 4o, benefitting from a large-scale architecture and optimization for diverse AI tasks.

In contrast, MedGemma 4B's lower relative performance underscores the impact of its smaller 4 billion parameter architecture, especially in complex tasks such as question-answering where deep contextual understanding is required. This suggests that, while MedGemma is efficient and resource-conscious, it presently lacks the sophistication required for some high-demand use cases.

The findings emphasize that model size and computational resources remain significant factors influencing AI performance, particularly for tasks demanding nuanced reasoning and domain expertise. However, MedGemma's competitive scores in classification and generation tasks indicate its potential value for applications with limited hardware or where faster inference is prioritized. Future work could explore model scaling, fine-tuning, or hybrid architectures to bridge this gap.

Overall, these results guide informed model selection based on specific use case requirements, balancing accuracy, domain sophistication, cost, and efficiency.

Recommendation:

Based on the evaluation results, it is recommended to select the AI model aligned with both task complexity and available computational resources. For critical applications demanding high accuracy, relevance, and domain appropriateness across varied tasks, GPT 4o is the preferred choice, offering the highest overall score at 99.1%. Its large-scale architecture ensures robust performance particularly for complex question-answering tasks requiring intricate contextual reasoning. Gemini 2.5 Pro is a close alternative, delivering similarly strong capabilities (97.2%) and may be favored when slightly lower infrastructure demands or comparative costs are factors. For scenarios prioritizing efficiency, lower latency, and resource constraints, MedGemma 4B remains a viable option, especially for classification and generation tasks where it maintains competitive scores despite its smaller 4-billion parameter size. Future use should weigh task requirements against resource availability, with potential to fine-tune MedGemma 4B or adopt hybrid approaches for enhanced performance without the overhead of trillion-parameter models.