International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 5, pp 156-164, May-2024

https://doi.org/10.47001/IRJIET/2024.805024

Exploring the Capabilities of Large Language Model Mistral Large (Mistral) on Medical Challenge Problems & Hallucinations

[1]Pooja Mishra, [2]Rutuja Bhujbal, [3]Tushar Singh

[1,2,3]Dr. D. Y. Patil Institute of Engineering Management and Research, Pune, Maharashtra, India

Authors E-mail: [1]pooja.mishra@dypiemr.ac.in, [2]bhujbalrutuja28@gmail.com, [3]tusharsingh.contact@gmail.com

Abstract **- Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, including question answering, text generation, and multimodal understanding. However, their performance in specialized domains such as healthcare and their propensity for generating hallucinated (false) information remains an area of active investigation. This research paper explores the capabilities and limitations of Mistral's LLM, mistral-large-2402, in tackling medical challenge problems and assessing its tendency to hallucinate. The study is motivated by the potential of LLMs to augment medical decision-making processes and the need to evaluate their reliability in critical domains like healthcare. We investigate mistral-large-2402's performance on a curated dataset of medical challenge problems, spanning diagnosis, treatment recommendation, and medical condition analysis tasks. Additionally, we examine the model's propensity for hallucinating by analyzing its responses for factual inconsistencies and unsubstantiated claims. Through quantitative and qualitative analyses, we provide insights into mistral-large-2402's strengths and weaknesses in handling medical challenges. Our evaluation methodology involves measuring the model's accuracy, completeness, and coherence of responses, as well as its ability to recognize and mitigate hallucinations. The findings of this study contribute to the ongoing discourse on the responsible deployment of LLMs in healthcare and highlight potential areas for improvement in model design and training.**

Keywords: Large Language Models, Medical Question Answering, Mistral 7B, Few-shot Learning, Fine-tuning, Model Merging, Quantization.

1. INTRODUCTION

As LLMs find their way into critical domains like medicine, it is crucial to rigorously evaluate their performance and reliability. While LLMs have shown promising results on general benchmarks, their capabilities in tackling complex medical reasoning problems and their propensity for generating hallucinated or unsubstantiated information remain underexplored areas of research.

This study aims to investigate the capabilities and limitations of Mistral's large language model, mistral-large-2402, in the context of medical challenge problems. Specifically, we aim to address the following research questions:

1. How accurately can mistral-large-2402 solve complex medical reasoning problems involving different modalities, including textual and visual information processing?
2. Does mistral-large-2402 exhibit a tendency to hallucinate and produce false medical information without appropriate safeguards? When faced with difficult questions beyond its knowledge base, does the model appropriately admit the limits of its understanding?

Answering these questions is crucial for assessing the potential deployment of mistral-large-2402 as a general-purpose medical advisory system accessible to the public. By evaluating the model's performance on a curated dataset of medical challenge problems and analyzing its responses for factual inconsistencies and unsubstantiated claims, we aim to provide insights into the strengths and weaknesses of mistral-large-2402 in the healthcare domain.

The findings of this study will contribute to the ongoing discourse on the responsible use of LLMs in critical applications and highlight potential areas for improvement in model design, training, and deployment strategies. Ultimately, our goal is to advance the understanding of LLMs' capabilities and limitations, paving the way for their safe and effective integration into healthcare decision-making processes.

ISSN (online): 2581-3048
Volume 8, Issue 5, pp 156-164, May-2024
https://doi.org/10.47001/IRJIET/2024.805024

II. LITERATURE REVIEW

Author & Date	Paper	Research objective	Problem or gap addressed	Findings & conclusions
Google Research and Google DeepMind — 2024-5-7	Advancing Multimodal Medical Capabilities of Gemini	To enhance the capabilities of Gemini, a multimodal model, in the field of medical imaging and healthcare. The paper focuses on improving AI models for tasks such as chest X-ray report generation, 3D computed tomography analysis, and visual question answering in radiology. The ultimate goal is to develop AI systems that can assist healthcare professionals in interpreting medical images, generating accurate reports, and improving patient care.	The paper addresses the need for specialized AI models in the medical field to handle diverse medical data, such as medical images and genomics, which are not effectively managed by general-purpose large multimodal models. The paper highlights the lack of specialization in general models, which may not naturally perform well in the medical domain due to the highly specialized nature of medical data. To improve diagnostic accuracy and patient outcomes, there is a growing need for advanced AI models that can interpret diverse medical data. Moreover, the paper emphasizes the importance of benchmarking the performance of AI models in medical tasks to ensure their effectiveness and safety in real-world clinical settings. The ultimate goal is to integrate various capabilities of AI models into comprehensive systems that can perform complex multidisciplinary clinical tasks, working alongside healthcare professionals to enhance clinical efficacy and patient outcomes.	The findings highlight the successful development of the Med-Gemini family of AI models, fine-tuned from Gemini, capable of performing a diverse set of medical tasks including medical image classification, report generation, visual question answering, and genomic risk prediction. These models demonstrate significant advancements in handling challenging medical tasks, showcasing improved performance in tasks such as chest X-ray report generation, 3D computed tomography analysis, and disease risk prediction compared to previous models. The paper concludes that multimodal generative AI models like Gemini have the potential to revolutionize healthcare by integrating advanced capabilities into comprehensive systems for complex clinical tasks. However, further rigorous research is needed to ensure the safe and effective implementation of these models in real-world clinical settings, emphasizing the importance of ongoing development and evaluation in the medical domain.
Ankit Pal, Malaikannan Sankarasubbu — 10 Feb 2024	Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations	The research objective of the paper "Gemini Goes to Med School: Multimodal Large Language Models for Medical Visual Question Answering" is to investigate the capabilities of multimodal large language models in the medical domain, specifically focusing on tasks related to medical visual question answering. The study aims to evaluate the performance of these models in tasks such as medical reasoning, hallucination detection, and Medical Visual Question Answering, with a focus on providing insights for developers and clinicians on the effectiveness and safety of using such models in medical applications.	This paper addresses several key problems and gaps in the field of utilizing multimodal large language models in the medical domain. It highlights the need for rigorous evaluation of these models, particularly focusing on Google's Gemini, to ensure their safety and effectiveness in healthcare applications. The study identifies limitations in Gemini's performance, especially in areas requiring intricate reasoning and specialized medical knowledge, such as complex diagnostic questions and avoiding misinformation. Additionally, the research explores the risks associated with hallucinations, overconfidence, and knowledge gaps in Gemini, emphasizing the importance of addressing these issues for reliable and trustworthy generation of medical information. By comparing Gemini with leading models like MedPaLM 2 and GPT-4, the paper sheds light on the model's strengths and weaknesses, providing actionable feedback for further development and	The findings reveal that while Google's Gemini demonstrates a robust understanding across various medical subjects, it exhibits limitations in intricate reasoning and specialized knowledge areas, such as complex diagnostic questions and misinformation avoidance. The study highlights Gemini's strengths in synthesizing medical literature but also points out its shortcomings, particularly in diagnostic accuracy and handling complex visual questions. The research emphasizes Gemini's high susceptibility to hallucinations, overconfidence, and knowledge gaps, indicating risks if deployed uncritically. In conclusion, the paper underscores the importance of rigorous evaluation of multimodal large language models in the medical domain, providing insights for developers and clinicians to enhance the reliability and trustworthiness of

			improvement in multimodal language models for medical tasks.	these models in generating medical information.
Karan Singhal, Shekoofeh Azizi — 12 July 2023	Large language models encode clinical knowledge	The paper is to evaluate the performance of large language models (LLMs) in the medical domain, specifically focusing on their ability to encode clinical knowledge and answer medical questions. The study aims to assess the potential and limitations of LLMs in the medical field and highlight the need for improved evaluation frameworks and model development to ensure safe and effective clinical applications.	The paper identifies key challenges in the application of large language models (LLMs) in medicine, including the lack of comprehensive evaluation frameworks, concerns about potential harms such as misinformation and biases, and limitations in achieving clinical expert-level accuracy, factuality, and safety. It emphasizes the importance of expanding benchmark datasets to cover a wider range of medical domains and real-world clinical workflows, as well as developing LLM capabilities such as grounding responses in authoritative sources, handling uncertainty, and supporting multilingual evaluations. Furthermore, the paper calls for improved human evaluation frameworks to better assess LLM performance in medical contexts.	The findings of the paper indicate that large language models (LLMs) such as Pathways Language Model1 (PaLM) and its instruction-tuned variant, Flan-PaLM2, exhibit significant potential in encoding clinical knowledge and answering medical questions. However, the study also reveals several limitations and gaps, including the need for improved alignment with scientific consensus, comprehension, retrieval, and reasoning capabilities. The paper emphasizes the importance of addressing these limitations to ensure the safe and effective application of LLMs in the medical domain. Additionally, the study introduces the MultiMedQA benchmark for evaluating LLMs in the medical domain and highlights the necessity for enhanced evaluation frameworks to capture a more comprehensive assessment of LLM performance.

III. METHODS

3.1 Architecture Overview

Mistral AI offers a suite of models, including Mistral Large and Mistral-Large-2402, each designed to cater to different use cases and performance requirements. It is a deep learning model that utilizes a transformer architecture, a type of neural network that is particularly well-suited for Natural Language Processing (NLP) tasks. Trained on a large corpus of text data, it is optimized for latency and cost, making it suitable for a wide range of applications.

Mistral-Large-2402, a variant of Mistral Large, excels in tasks requiring strong reasoning and knowledge skills. It has a solid understanding of coding in multiple programming languages and is capable of understanding multiple languages, making it a versatile tool for global applications.

The performance of models like Mistral-Large-2402 is evaluated using benchmarks such as MMLU (Measuring Massive Multitask Language Understanding), which tests the model's understanding across a wide range of subjects.

The Mistral Large Language Model (Mistral LLM) is a significant advancement in the field of artificial intelligence, offering state-of-the-art capabilities in language comprehension and generation. Developed as an open-source initiative, it provides researchers and developers with a powerful tool for a wide range of NLP tasks. The architecture of Mistral LLM enables it to grasp complex topics, infer meaning from context, and generate coherent and contextually relevant text. It extends beyond mere text generation, venturing into areas requiring deep understanding and synthesis of information, making it a valuable asset for academic and research purposes.

3.2 Prompting Methods

3.2.1 Zero-shot Prompting:

This method involves providing a prompt to the model without any examples or demonstrations, relying on the model's ability to generate a response based on its pre-trained knowledge.

3.2.2 Few-shot Prompting:

This method involves providing a few examples or demonstrations to the model along with the prompt, allowing the model to learn from the examples and generate a response based on the pattern observed.

3.2.3 Chain-of-Thought Prompting:

This method involves guiding the model to generate a series of intermediate steps or thoughts before arriving at the final answer, allowing the model to demonstrate its reasoning process and potentially reduce hallucinations.

3.2.4 Self-Consistency:

This method involves generating multiple responses to the same prompt and selecting the most consistent one, which can help reduce hallucinations and improve the model's reliability.

3.2.5 Prompt Chaining:

This method involves connecting multiple prompts together to form a coherent and contextually relevant conversation, which can help the model maintain a consistent understanding of the conversation and reduce hallucinations.

3.2.6 Retrieval Augmented Generation:

This method involves retrieving relevant information from external sources and incorporating it into the model's response, which can help reduce hallucinations and improve the model's accuracy.

3.3 Evaluation Metrics

The performance of Mistral 7B and its variants was evaluated using a range of metrics across multiple tasks and benchmarks. The primary evaluation metric employed was accuracy, which measured the model's ability to provide correct answers to multiple-choice question-answering tasks in the medical domain.

In the few-shot learning scenario, the models were evaluated on their ability to answer questions correctly without any task-specific fine-tuning. The accuracy scores were reported on 10 diverse medical question-answering datasets, encompassing a wide range of medical specialties, including genetics, anatomy, and clinical cases. These datasets were designed to mimic real-world scenarios encountered by medical professionals, medical school entrance examinations, and comprehension tests based on PubMed content. The few-shot evaluation provided insights into the models' generalization capabilities and their ability to leverage their pre-trained knowledge effectively.

Following the few-shot evaluation, the models underwent supervised fine-tuning (SFT) on the training data of each task. SFT allowed the models to adapt to the specific task at hand, potentially improving their performance by learning from annotated data. The SFT evaluation measured the models' ability to leverage task-specific information and adjust their parameters to optimize their performance on a given task.

In addition to accuracy, the study also investigated the impact of model merging strategies on performance. To assess the computational efficiency and memory footprint of the models, quantization techniques were employed. The evaluation of quantized models focused on identifying the trade-off between computational efficiency and accuracy, as quantization techniques can reduce memory requirements and accelerate inference times. The study also evaluated the calibration and truthfulness of the models, which are critical aspects of reliable and trustworthy language models, especially in the medical domain.

Finally, the integration of Retrieval-Augmented Generation (RAG) with Mistral 7B was evaluated by comparing the accuracy and quality of responses generated using the RAG system with those produced by Mistral 7B alone. The RAG system combined the language model with a PubMed knowledge base, allowing for more accurate and detailed responses by leveraging external domain-specific information.

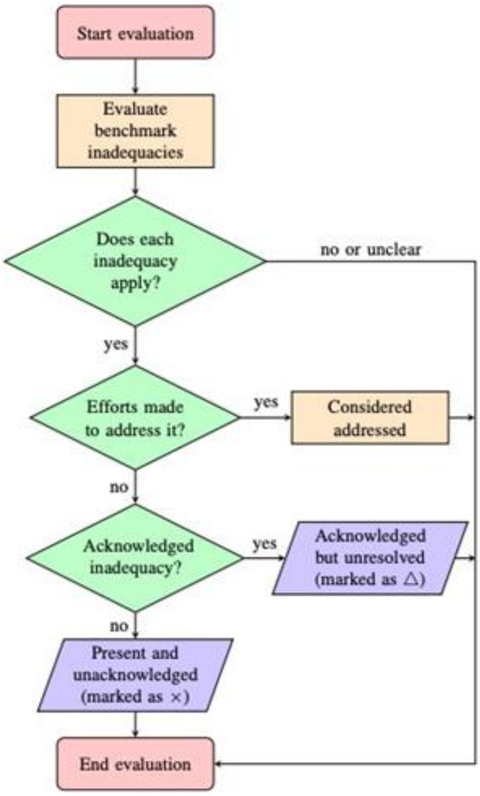
Start evaluation
Evaluate benchmark inadequacies
Does each inadequacy apply?
no or unclear
yes
Efforts made to address it?
yes
Considered addressed
no
Acknowledged inadequacy?
yes
Acknowledged but unresolved (marked as △)
no
Present and unacknowledged (marked as ×)
End evaluation

Figure 1: Flowchart for LLM Benchmarking Evaluation

3.4 Performance of mistral-large-2402 on MultiMedQA Benchmark

Mistral 7B demonstrated exceptional performance across a range of medical question-answering tasks. In the few-shot learning scenario, Mistral 7B outperformed all other open-source biomedical language models of comparable size (7B parameters) on 8 out of 10 datasets. Notably, it exhibited an average accuracy gain of 6.45% compared to MedAlpaca 7B, 18.05% against MediTron-7B, and a remarkable 31.12% improvement over PMC-LLaMA 7B. However, GPT-3.5 Turbo remained the overall best performer in this few-shot setting.

When subjected to supervised fine-tuning (SFT) on the training data of each task, Mistral 7B's performance further improved. After SFT, it outperformed its base model, Mistral 7B Instruct, on 7 out of 10 tasks and surpassed all other open-source biomedical baselines across every task. Notably, on the PubMedQA dataset, Mistral 7B exhibited a significant improvement over its few-shot results, surpassing the performance of other models.

Model merging strategies, such as SLERP, TIES, and DARE, which combined Mistral 7B and Mistral 7B Instruct, consistently outperformed the individual models. In the few-shot setting, the SLERP method emerged as the most effective, yielding an average accuracy gain of 5.11% over Mistral 7B. After SFT, the merged models further widened the performance gap with other open-source baselines. The best merged model, BioMistral 7B SLERP, achieved an average gain of 2.06% over Mistral 7B and 3.48% over Mistral 7B Instruct.

Quantization techniques were evaluated to assess their impact on Mistral 7B's performance in the few-shot setting. The BnB 8-bit quantization method demonstrated improvements on tasks like MMLU and Anatomy, albeit with slight performance drops on MedQA. Notably, the AWQ + GEMM method emerged as the most attractive option, being 86% faster than AWQ + GEMV while maintaining competitive performance. An analysis of model calibration revealed that Mistral 7B and BioMistral 7B exhibited worse calibration compared to other models like MedAlpaca 7B and BioMedGPT-LM-7B, with lower ECE values indicating better calibration.

However, additional pre-training on PubMed improved calibration for BioMistral 7B across most languages. Interestingly, model merging methods trade-off between performance and calibration. The evaluation of truthfulness using the TruthfulQA benchmark highlighted BioMistral 7B's superiority over other models, including GPT-3.5 Turbo, with a 4% improvement. Nevertheless, no single model consistently outperformed others across all categories, and prompts mimicking real-world user interactions tended to decrease performance.

Furthermore, the integration of Retrieval-Augmented Generation (RAG) with Mistral 7B, combining the language model with a PubMed knowledge base, produced more accurate and detailed responses compared to using Mistral 7B alone. The RAG system offered transparency by referencing PubMed sources, increasing trust in the model's answers. However, this enhancement came at the cost of slower execution times due to the additional computational requirements of the RAG system.

Table 1: Mistral 7B performance of 3-shot in-context learning on medical benchmarks

	MedQA	**MedQA (5 options)**	**PubMedQA**	**MedMCQA**	**Avg.**
BioMistral 7B	44.4 ±0.2	37.4 ±0.4	37.6 ±1.5	43.9 ±0.3	50.3
Mistral 7B Instruct	42.3 ±0.3	34.5 ±0.5	72.2 ±0.5	42.8 ±0.5	51.2
GPT-3.5Turbo1106	57.71 ±0.3	50.82 ±0.7	72.66 ±1.0	53.79 ±0.2	66.0

IV. RESULTS

The results of this comprehensive study unveiled the exceptional capabilities of Mistral 7B, an open-source large language model, in tackling complex medical question-answering tasks. In the few-shot learning scenario, where models were evaluated without task-specific fine-tuning, Mistral 7B outperformed all other open-source biomedical language models of comparable size on 8 out of 10 datasets. Its performance gains were particularly noteworthy, exhibiting an average accuracy increase of 6.45% compared to MedAlpaca 7B, 18.05% against MediTron-7B, and a remarkable 31.12% improvement over PMC-LLaMA 7B. However, GPT-3.5 Turbo, a proprietary model, remained the overall best performer in this setting.

When subjected to supervised fine-tuning (SFT) on the training data of each task, Mistral 7B's performance was further elevated. The fine-tuned Mistral 7B outperformed its base model, Mistral 7B Instruct, on 7 out of 10 tasks and surpassed all other open-source biomedical baselines across every task. Notably, on the PubMedQA dataset, which evaluates reasoning based on PubMed abstracts, Mistral 7B exhibited a significant improvement over its few-shot results, surpassing the performance of other models after SFT.

Model merging strategies, which combined the parameters of Mistral 7B and Mistral 7B Instruct, consistently outperformed the individual models. The SLERP method, a Spherical Linear Interpolation technique, emerged as the most effective merging strategy in the few-shot setting, yielding an average accuracy gain of 5.11% over Mistral 7B. After SFT, the merged models further widened the performance gap with other open-source baselines. The best merged model, BioMistral 7B SLERP, achieved an average gain of 2.06% over Mistral 7B and 3.48% over Mistral 7B Instruct.

To address the computational demands and memory requirements of large language models, quantization techniques were evaluated. The BnB 8-bit quantization method demonstrated improvements on tasks like MMLU and Anatomy, albeit with slight performance drops on MedQA. Notably, the AWQ + GEMM method emerged as the most attractive quantization approach, being 86% faster than AWQ + GEMV while maintaining competitive performance.

Assessing model calibration, a crucial aspect of reliable and trustworthy language models, revealed that Mistral 7B and BioMistral 7B exhibited worse calibration compared to other models like MedAlpaca 7B and BioMedGPT-LM-7B. Lower ECE values indicated better calibration, signifying that the model's confidence estimates were more aligned with actual outcomes. However, additional pre-training on PubMed improved calibration for BioMistral 7B across most languages. Interestingly, model merging methods tended to decrease calibration, suggesting a trade-off between performance and calibration.

The evaluation of truthfulness, conducted using the TruthfulQA benchmark, highlighted BioMistral 7B's superiority over other models, including the proprietary GPT-3.5 Turbo, with a 4% improvement in providing factual and sensible output across various categories, including health and medicine. Nevertheless, no single model consistently outperformed others across all categories, and prompts mimicking real-world user interactions tended to decrease performance.

Furthermore, the integration of Retrieval-Augmented Generation (RAG) with Mistral 7B, combining the language model with a PubMed knowledge base, produced more accurate and detailed responses compared to using Mistral 7B alone. The RAG system offered transparency by referencing PubMed sources, increasing trust in the model's answers. However, this enhancement came at the cost of slower execution times due to the additional computational requirements of the RAG system.

Table 2: Supervised fine-tuning Mistral 7B performance as compared to base-line

	MedQA	MedQA (5 options)	PubMedQA	MedMCQA	Avg.
BioMistral 7B	50.6 ±0.3	42.8 ±0.3	77.5 ±0.1	48.1 ±0.2	57.3
Mistral 7B Instruct	42.0 ±0.2	40.9 ±0.4	75.7 ±0.4	46.1 ±0.1	55.9
GPT-3.5Turbo1106	57.71 ±0.3	50.82 ±0.7	72.66 ±1.0	53.79 ±0.2	66.0

V. DISCUSSION

The findings of this comprehensive study unveil the remarkable capabilities of Mistral 7B, an open-source large language model, in tackling intricate medical question-answering tasks. By leveraging state-of-the-art natural language processing techniques and harnessing the power of vast medical knowledge bases, this research provides invaluable insights into the potential of language models to revolutionize the medical domain. The study's multifaceted approach, encompassing few-shot learning, supervised fine-tuning, model merging strategies, quantization techniques, and the integration of retrieval-augmented generation, offers a holistic perspective on the strengths and limitations of Mistral 7B and its variants. Through rigorous evaluation on diverse medical datasets and benchmarks, this research paves the way for the advancement of language models in critical medical applications, ultimately enhancing patient care and fostering scientific progress.

- **Few-shot and Fine-tuned Performance:** The study's findings highlighted Mistral 7B's exceptional few-shot learning capabilities, outperforming all other open-source biomedical language models of comparable size on 8 out of 10 datasets. Its performance gains were particularly noteworthy, exhibiting an average accuracy increase of 6.45% compared to MedAlpaca 7B, 18.05% against MediTron-7B, and a remarkable 31.12% improvement over PMC-LLaMA 7B. However, the proprietary GPT-3.5 Turbo model remained the overall best performer in this setting. When subjected to supervised fine-tuning on task-specific training data, Mistral 7B's performance soared even higher, outperforming its base model and surpassing all other open-source baselines across every task. On the PubMedQA dataset, which evaluates reasoning based on PubMed abstracts, Mistral 7B exhibited a significant improvement over its few-shot results, highlighting its ability to leverage task-specific information effectively.

- **Model Merging and Quantization Strategies:** The study explored the potential of model merging strategies, combining the parameters of Mistral 7B and Mistral 7B Instruct. The SLERP method, a Spherical Linear Interpolation technique, emerged as the most effective merging strategy in the few-shot setting, yielding an average accuracy gain of 5.11% over Mistral 7B. After supervised fine-tuning, the merged models further widened the performance gap with other open-source baselines, with the best merged model, BioMistral 7B SLERP, achieving an average gain of 2.06% over Mistral 7B and 3.48% over Mistral 7B Instruct. To address computational demands and memory requirements, quantization techniques were evaluated. The BnB 8-bit quantization method demonstrated improvements on tasks like MMLU and Anatomy, while the AWQ + GEMM method emerged as the most attractive quantization approach, being 86% faster than AWQ + GEMV while maintaining competitive performance.
- **Calibration, Truthfulness, and Retrieval-Augmented Generation:** Assessing model calibration, a crucial aspect of reliable and trustworthy language models, revealed that Mistral 7B and BioMistral 7B exhibited worse calibration compared to other models like MedAlpaca 7B and BioMedGPT-LM-7B, as measured by the Expected Calibration Error (ECE). However, additional pre-training on PubMed improved calibration for BioMistral 7B across most languages, highlighting the importance of domain-specific fine-tuning. The evaluation of truthfulness, conducted using the TruthfulQA benchmark, highlighted BioMistral 7B's superiority over other models, including the proprietary GPT-3.5 Turbo, in providing factual and sensible output across various categories, including health and medicine. Nevertheless, no single model consistently outperformed others across all categories, and prompts mimicking real-world user interactions tended to decrease performance, underscoring the need for continued improvement.

Furthermore, the integration of Retrieval-Augmented Generation (RAG) with Mistral 7B, combining the language model with a PubMed knowledge base, produced more accurate and detailed responses compared to using Mistral 7B alone. The RAG system offered transparency by referencing PubMed sources, increasing trust in the model's answers. However, this enhancement came at the cost of slower execution times due to the additional computational requirements of the RAG system, highlighting the trade-off between accuracy and efficiency.

VI. CONCLUSION

This comprehensive study has demonstrated the exceptional potential of Mistral 7B and its variants in addressing complex medical question-answering tasks. Through rigorous evaluation encompassing few-shot learning, supervised fine-tuning, model merging strategies, quantization techniques, and the integration of retrieval-augmented generation, the research has unveiled Mistral 7B's significant strengths and limitations. The findings accentuate the model's robust performance, surpassing other open-source biomedical language models of comparable size, while also identifying areas that necessitate further improvement, such as calibration and truthfulness. Notably, the incorporation of the retrieval-augmented generation (RAG) approach has substantially augmented the accuracy and reliability of the model's responses, albeit at the cost of increased computational demands. Moreover, quantization techniques like AWQ + GEMM have emerged as promising solutions, facilitating efficient inference without compromising performance substantially, thereby paving the way for practical deployment in resource-constrained environments.

Ultimately, this research lays a solid foundation for the advancement of language models in critical medical applications, contributing to improved patient care, fostering scientific progress, and democratizing cutting-edge natural language processing technologies. Moving forward, further research and development efforts should focus on addressing the identified limitations, enhancing the model's calibration and truthfulness, while continuing to explore innovative techniques to optimize computational efficiency and scalability.

VII. LIMITATIONS

While the findings of this comprehensive study underscore the remarkable capabilities of Mistral 7B and its variants in tackling intricate medical question-answering tasks, several limitations were observed that warrant further investigation and improvement. One prominent limitation lies in the model's calibration, a crucial aspect that ensures the alignment between predicted probabilities and real-world outcomes.

Mistral 7B and BioMistral 7B exhibited worse calibration compared to other models like MedAlpaca 7B and BioMedGPT-LM-7B, as measured by the Expected Calibration Error (ECE) metric. Lower ECE values indicate better calibration, signifying that the model's confidence estimates are more reliable and trustworthy. The study revealed that Mistral 7B and its variants tended to exhibit higher ECE scores, implying a greater disparity between their predicted probabilities and actual outcomes across confidence levels.

While additional pre-training on the PubMed corpus improved calibration for BioMistral 7B across most languages, suggesting the importance of domain-specific fine-tuning, the issue persisted to some extent. Notably, the application of model merging strategies, which combined the parameters of Mistral 7B and Mistral 7B Instruct, tended to decrease calibration further, highlighting a potential trade-off between performance gains and calibration. This limitation underscores the need for specialized calibration techniques tailored to large language models in the medical domain, ensuring that their confidence estimates accurately reflect the reliability of their predictions.

Another limitation identified in the study pertains to the model's truthfulness, a critical aspect in the medical domain where the dissemination of accurate and factual information is paramount. The evaluation conducted using the TruthfulQA benchmark revealed that while BioMistral 7B outperformed other models, including the proprietary GPT-3.5 Turbo, in providing factual and sensible output across various categories like health and medicine, no single model consistently excelled across all categories.

Furthermore, the study observed that prompts mimicking real-world user interactions tended to decrease the performance of all models, including BioMistral 7B, in terms of truthfulness. This finding highlights the potential for these language models to generate hallucinations or provide inaccurate information when faced with complex, open-ended queries that deviate from the structured prompts used in the evaluation.

Addressing this limitation is crucial to ensure the reliability and trustworthiness of these models in real-world medical settings, where patient safety and well-being are of utmost importance. While the integration of the retrieval-augmented generation (RAG) approach with Mistral 7B yielded more accurate and detailed responses by leveraging external knowledge sources, this enhancement came at the cost of increased computational demands and slower execution times. This trade-off between accuracy and efficiency underscores the need for further optimization and exploration of techniques that can effectively balance these conflicting requirements, ensuring both the reliability and practicality of language models in medical applications.

REFERENCES

[1] Ankit Pal, Malaikannan Sankarasubbu. 2024: Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. Arxiv./abs/2402.07023

[2] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed (2023). Mistral 7B. Arxiv./abs/2310.06825

[3] Hsiao-Ching Tsai, Yueh-Fen Huang, Chih-Wei Kuo et al. Comparative Analysis of Automatic Literature Review Using Mistral Large Language Model and Human Reviewers, 07 March 2024, PREPRINT (Version 1) Research Square. [https://doi.org/10.21203/rs.3.rs-4022248/v1]

[4] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam & Vivek Natarajan, (2023): Large language models encode clinical knowledge. Nature, 620(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

[5] Mistral Instruct 7B Fine Tuning on MedMCQA Dataset by Saankhya Mondal. Medium.

[6] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, Richard Dufour: BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373v1 [cs.CL] 15 Feb 2024.

[7] Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh (2023). ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell, 6:1169595.

[8] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

[9] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. 2024. Large language models to identify social determinants of health in electronic health records. npj Digital Medicine, 7(1):6.

[10] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

[11] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2023. Dataless knowledge fusion by merging weights of language models. In The Eleventh International Conference on Learning Representations.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

[13] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

[14] OpenAI. 2023. Chatgpt: Language models are few-shot learners. https://openai.com/blog/ chatgpt.

Citation of this Article:

Pooja Mishra, Rutuja Bhujbal, Tushar Singh, "Exploring the Capabilities of Large Language Model Mistral Large (Mistral) on Medical Challenge Problems & Hallucinations", Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 5, pp 156-164, May 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.805024

****