Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 8, pp. 1713–1722
DOI : 10.47974/JSMS-1403

Comparative analysis of medical knowledge and inferential capabilities of modern LLMs

Muhammad Badruddin Khan *
Abdul Khader Jilani Saudagar †
Information Systems Department
College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh 11432
Saudi Arabia

Abstract

This paper presents an evaluation of four large language models (LLMs)—GPT-4o, Claude-3.5-Sonnet, Gemini-1.5-Pro, and Llama-3.1-405B-T based on word problems that were formulated to check LLMs level of "understanding" of medical knowledge and their "inferential capabilities". We constructed a small synthetic dataset by providing pattern to LLM and then asking LLM to generate few word problems related to heartbeat, and evaluated responses of different LLMs for each word problem. Models were evaluated based on their robustness to changing scenarios. Moreover, their responses were compared in order to understand their level of learning about medical knowledge. Consistency of models was checked by providing them answer choices in random order. Then models' answers were compared with their previous solutions. It was found that all the LLMs involved in the study showed inconsistent behavior. The study also reveals that Gemini and LLama did not show required resilience to changing input. GPT-4o and Claude gave similar answers for most of the queries and showed robustness to changing inputs. The study reveals very interesting insights about strengths and weaknesses of each of LLM and can be very helpful for healthcare professionals in deciding level of adoption of modern artificial intelligence (AI) technologies for identification of heartbeat related disorders like Arrhythmia.

Subject Classification: *Primary 00A05, Secondary 68T50.*

Keywords: *Large language models (LLMs), AI technologies, Healthcare, Reasoning, Heartbeat, Arrhythmia*

* *E-mail:* **mbkhan@imamu.edu.sa (Corresponding author)**
† *E-mail:* **aksaudagar@imamu.edu.sa**

1. Introduction

To enable machines to understand and generate human-like text is one of the greatest achievements of modern artificial intelligence (AI) endeavors. The advent of internet and social media enabled the availability of a massive amount of text (like books, reviews, posts, conversations, and opinions) and other types of data that were used to train machines to learn patterns in the given modality. Generative AI used textual form of data to make large language models (LLMs) learn to perform multiple generative tasks like story writing, languages translation, and ideas brainstorming. The natural communication between humans and machines meant that the technology is available to be used by different segments of life. One of the areas where LLMs can assist humanity is to help in improving healthcare.

LLMs are trained with huge amount of medical data too. The medical knowledge of modern LLMs is unprecedented and amazing. LLMs use this medical knowledge in their reasoning process. The reasoning process is available to a viewer and most of the times, it is very impressive to a layman. LLMs can be helpful in analysis of medical records. They can assist doctors in diagnoses of disease. They can be used by hospitals to answer patients' queries. Thus, they can support health professionals in accomplishing their tasks thus saving precious resources like time and money.

However, unlike other areas, healthcare area is very sensitive in the sense that it deals with human life. A slight mistake can result in huge disaster for a patient and his/her relatives. Hence, inclusion of LLMs in health care activities demand highest level of accuracy. Intentional or accidental LLMs vulnerability exploitation is a serious risk that is of the highest concern of any healthcare professional. One of the life-threatening conditions is the heart-attack scenario. For example, **Arrhythmia** is a condition where the heart beats irregularly—too fast, too slow, or with unevenly rhythm. The consequences can range from feeling of dizziness and faintness to heart, brain or other organ damage leading to life threatening strokes, heart failure or cardiac arrest.

Large Language Models (LLMs) can be helpful in arrhythmia management by analyzing patient data, summarizing their medical records, and assisting doctors in interpreting ECG results. Not only doctors, but they can also help patients by explaining to them their health conditions and answering their common queries. Descriptive analytics use different statistical measures to summarize the situation. In case of

arrhythmia patient, it can be average of heartbeats recorded at different times. Another measure is range that tells doctor about the highest level of fluctuations observed in heartbeat rates.

In this paper, we present our work related to evaluation of LLMs 'understanding' of variation of heartbeat rates due to change of human state. The work is significant because it can help us in establishing whether we can trust LLMs' verdicts when we provide them with patients' data in form of natural language. The paper provides information about related research in the literature review section (section 2). The section is followed by section 3 discussing our methodology. The most important section 4 comprises of the results and discussion of our experiments. The final section of paper (section 5) i.e., conclusion summarizes the work and discusses future areas for improvement.

2. Literature Review

This research work is inspired by recent developments in assessment and evaluation of the mathematical reasoning capabilities of Large Language Models (LLMs). Key findings from [1] revealed significant limitations of LLMs in various scenarios, such as minor variations in numerical values resulting in performance discrepancies, sensitivity to rising question complexity, and susceptibility to irrelevant information. Work of [1] is not specific to particular domain. Unlike this work is related to very specific domain of healthcare. The paper [1] used GSM8K benchmark which was introduced by [2] and is a dataset of 8500 high quality linguistically diverse grade school math word problems whereas current study formulated its own questions i.e., word problems after lot of experimentations. [3] investigated the limits of Transformer large language models (LLMs) across three representative compositional tasks -- multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. Their work was based on the researchers' observations that on some surprisingly trivial problems, models fail.

"Are such error-prone, hallucinating models suitable to be introduced in healthcare domain?" becomes a very important question to answer. With all limitations, the integration of Large Language Models (LLMs) in healthcare industry is advancing and has attracted significant attention of many researchers due to their potential to transform various aspects of healthcare domain. However, before discussing these studies, we will introduce four LLMs that were used in this study for evaluation and comparison purpose.

GPT-4.0 [4] is a large language model developed by OpenAI and was released in March 2023. It is known for its improved reasoning capabilities and multimodal functionality. GPT-4o ("o" for "omni") [5] was released in May 2024 and is a multilingual, multimodal generative pre-trained transformer by OpenAI.

Claude-3.5-Sonnet [6] is a large language model from Anthropic's Claude series, which emphasizes safety and alignment through "Constitutional AI.". According to [6], "Claude 3.5 Sonnet sets new industry benchmarks for graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval)."

Gemini [7] is a LLM developed by Google DeepMind, and build its language understanding and advanced reasoning capabilities based on Google's expertise in different areas of machine learning. According to [8], "1.5 pro achieves near-perfect recall on long-context retrieval tasks across modalities, unlocking the ability to accurately process large-scale documents, thousands of lines of code, hours of audio, video, and more."

LLama (Large Language Model Meta AI) is a series of open-source large language models developed by Meta (formerly Facebook) [9]. In 2024, Meta released a collection of large AI models, including Llama 3.1 405B [10], that has performance comparable to the most advanced closed-source models. It is also important to mention that all experiments for the purpose of this study, were performed in "poe" environment [11] where interaction with all these four models is possible.

Before introducing our methodology, we will mention few studies that discusses the role of LLMs in healthcare industry. For example [12] highlights the potential of ChatGPT in improving healthcare education, research, and practice. Similarly [13] evaluated ChatGPT's feasibility across clinical and research scenarios, and pointed out that "it is important to recognize and promote education on the appropriate use and potential pitfalls of AI-based LLMs in medicine." ChatDoctor in [14] is a proof of significant advancement in medical LLMs, and demonstrated a significant improvement in understanding patient inquiries and providing accurate advice after large language model meta-AI (LLaMA) was fine-tuned using 100,000 patient-doctor dialogues.

3. Methodology

In this section, we will briefly provide an overview of the methodology. Following steps were followed for the purpose of the evaluation of different LLMs:

1. Identification of areas of consistency check (evaluation of ability of LLMs to 'understand' variability of heartbeat rate in different states)
2. Identification of areas of medical knowledge related to heartbeat rates
3. Formulation of word problems for identified area using medical knowledge provided by LLM and verified by domain experts (Many types of patterns were made, but in this paper, we will present one pattern for which few word problems were tested)
4. Querying four selected LLMs by giving word problems.
5. Analyze the response of LLMs.
6. Compilation of results for all word problems
7. Comparison of different LLMs performances.

The pattern question that was formulated to check LLM medical knowledge and its reasoning is given in Figure 1:

{Name of person} was sitting with a heart rate of {$40 \leq x \leq 80$} bpm. He suddenly starts {walking /running} due to fear. What will likely happen to his heart rate immediately after stopping?

A. It will stay elevated between 100 bpm and 110 bpm for some time.

B. it will stay elevated between 110 bpm and 120 bpm for some time.

C. It will stay elevated between 80 bpm and 90 bpm for some time.

D. It will stay elevated between 90 bpm and 100 bpm for some time.

Figure 1

Example of pattern used to generate multiple word problems

GPT-4o was asked to make word problems for only one name and it produced ten word-problems with initial heartbeat rates of 40, 50, 60, 70 and 80 bpm.

4. Results and discussion

The generated word problems were given as an input to different LLMs. Table 1 provide different responses of LLMs for ten word-problems.

It should be noted that for all tables, for GPT-4o, we have provided the heartbeat range along with answer as an example so that reader can understand which range is mapped with each answer. It can be seen in table 1 that for the 10 problems, GPT-4o and Claude Claude-3.5-Sonnet

Table 1

Responses of four LLMs for ten word-problems

From sitting state to:	Rest HB rate	GPT-4o	Claude	Gemini	LLama
1 (Walking)	40 bpm	C. 80–90 bpm	C	C	D
2 (Walking)	50 bpm	C. 80–90 bpm	C	C	D
3 (Walking)	60 bpm	D. 90–100 bpm	D	C	D
4 (Walking)	70 bpm	D. 90–100 bpm	D	C	D
5 (Walking)	80 bpm	A. 100–110 bpm	A	C	D
6 (Running)	40 bpm	A. 100–110 bpm	A	A /D	A
7 (Running)	50 bpm	A. 100–110 bpm	A	A /D	A
8 (Running)	60 bpm	B. 110–120 bpm	B	A /D	A
9 (Running)	70 bpm	B. 110–120 bpm	B	A /D	A
10(Running)	80 bpm	B. 110–120 bpm	B	A /D	A

provided similar answers whereas Gemini 1.5 pro and Llama 3.1 405B provided constant answers for scenario of walking and running without taking care of initial heartbeat rate (rest HB rate). One can read the reasoning to check the proficiency level of each of LLM.

In the second experiment, two ranges (E: 70-80 bpm and F: 120-130 bpm) were added in addition to four previous choices in order to check LLMs adaptability to new choices. Table 2 summarizes answers of the four LLMs.

Table 2

Responses of four LLMs for ten word-problems with 6 choices

From sitting state to:	Rest HB rate	GPT-4o	Claude	Gemini	LLama
1 (Walking)	40 bpm	E (70–80 bpm)	C	C	E
2 (Walking)	50 bpm	C (80–90 bpm)	C	C	E
3 (Walking)	60 bpm	D (90–100 bpm)	D	C	D
4 (Walking)	70 bpm	D (90–100 bpm)	D	C	D
5 (Walking)	80 bpm	A (100–110 bpm)	A	C	D
6 (Running)	40 bpm	A (100–110 bpm)	B	A /D	B
7 (Running)	50 bpm	A (100–110 bpm)	B	A /D	B
8 (Running)	60 bpm	B (110–120 bpm)	F	A /D	B
9 (Running)	70 bpm	B (110–120 bpm)	F	A /D	B
10(Running)	80 bpm	F (120–130 bpm)	F	A /D	B

Table 3

Responses of four LLMs for ten word-problems with 6 randomized choices

From sitting state to:	Rest HB rate	GPT-4o	Claude	Gemini	LLama
1 (Walking)	40 bpm	E (70–80 bpm)	C	C	E
2 (Walking)	50 bpm	C (80–90 bpm)	**D**	C	**D**
3 (Walking)	60 bpm	D (90–100 bpm)	D	C	D
4 (Walking)	70 bpm	D (90–100 bpm)	D	C	D
5 (Walking)	80 bpm	A (100–110 bpm)	A	C	D
6 (Running)	40 bpm	A (100–110 bpm)	B	A	**F**
7 (Running)	50 bpm	B (110–120 bpm)	B	**D**	B
8 (Running)	60 bpm	B (110–120 bpm)	F	A	B
9 (Running)	70 bpm	B (110–120 bpm)	F	A	B
10 (Running)	80 bpm	F (120–130 bpm)	F	A	B

Again, it can be seen in table 2 that Gemini was unable to take care of initial heartbeat rates. However, Llama showed some resilience for walking state related questions. After adding two more choices, disagreement between GPT-4o and Claude appeared for few questions. It should be noted that the questions were asked in same chat, hence, the memory content are also expected to make impact on LLMs' responses. Finally, In the third experiment, 6 choices from previous experiment were randomized to check LLMs responses. Gemini in this experiment was asked to give **only one** answer. Table 3 summarizes answers of the four LLMs for the third experiment.

In order to easily compare results of third experiment with the results of the second experiment, table 3 choice alphabets (A, B, C, D, E and F) were mapped with choice alphabets of table 2. It can be seen that all four LLMs provide inconsistent answer (for at least one question) when another arrangement of same choices were given to them in word problem. The inconsistent answers are highlighted by making them appeared in bold and underlined format. GPT-4o, Claude and Gemini had one inconsistent answer whereas Llama gave two inconsistent answers.

Three experiments results reveals that four LLMs can differ in their answers and all are prone to inconsistency when arrangement of answer choices is changed. This revelation can be helpful in improving performance of LLMs in their future versions as well as it can assist healthcare

professionals to decide the scope of adoption in their decision making. It is possible that with wrong 'thought' process, LLM accidentally arrives at correct answer hence this aspect should not be ignored.

5. Conclusion

This paper presents experiments' results for few simple word problems that were related to heartbeat. The pattern based on which the word problems were derived was very simple and the word problems were easy for LLM to understand. However, based on their learning level of medical knowledge, LLM gave different answers for different scenarios. It was also found that choices arrangements can affect LLM answers thus leading to inconsistent responses. As a future work, more complex patterns for different domains should be formulated and should be made public so that relevant professionals can use these patterns to generate word problems and observe LLMs' performance on them. In this study, we discussed the consistency in answers, but evaluation of reasoning of LLMs can help in determining the correctness of "though" process of LLMs.

Funding

The King Salman center For Disability Research provided funding for this work through Research Group no KSRG-2023-348.

Data Availability Statement

The data presented in this study are available upon request from the corresponding author.

Acknowledgement

The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2023-348.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," Oct. 07 (2024), arXiv: arXiv:2410.05229. doi: 10.48550/arXiv.2410.05229.

[2] K. Cobbe, A. S. Patel, J. S. Brown, and L. D. Miller, "Training verifiers to solve math word problems," arXiv, Nov. 18 (2021), arXiv:2110.14168. doi: 10.48550/arXiv.2110.14168.

[3] N. Dziri, S. J. Lee, P. M. Sharma, and R. T. Zhang, "Faith and fate: Limits of transformers on compositionality," arXiv, Oct. 31 (2023), arXiv:2305.18654. doi: 10.48550/arXiv.2305.18654.

[4] OpenAI, "GPT-4 technical report," arXiv, Mar. 4 (2024), arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[5] "GPT-4o," Wikipedia. Accessed: Dec. 15 (2024). [Online]. Available: https://en.wikipedia.org/w/index.php?title=GPT-4o&oldid=1262551536

[6] "Introducing Claude 3.5 Sonnet \ Anthropic." Accessed: Dec. 15 (2024). [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[7] "Gemini (language model)," Wikipedia. Accessed: Dec. 15 (2024). [Online]. Available: https://en.wikipedia.org/w/index.php?title=Gemini_(language_model)&oldid=1263935127

[8] "Gemini Pro - Google DeepMind." Accessed: Dec. 15 (2024). [Online]. Available: https://deepmind.google/technologies/gemini/pro/

[9] "Llama (language model)," Wikipedia. Accessed: Dec. 15 (2024). [Online]. Available: https://en.wikipedia.org/w/index.php?title=Llama_(language_model)&oldid=1264032930

[10] "Open-source artificial intelligence," Wikipedia. Accessed: Dec. 15 (2024). [Online]. Available: https://en.wikipedia.org/w/index.php?title=Open-source_artificial_intelligence&oldid=1263322594

[11] "Quora," Wikipedia. Accessed: Dec. 15 (2024). [Online]. Available: https://en.wikipedia.org/w/index.php?title=Quora&oldid=1262150310#Poe

[12] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid

Concerns," *Healthc. Basel Switz.*, vol. 11, no. 6, p. 887, Mar. (2023), doi: 10.3390/healthcare11060887.

[13] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios," *J. Med. Syst.*, vol. 47, no. 1, p. 33 (2023), doi: 10.1007/s10916-023-01925-4.

[14] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," Jun. 24 (2023), arXiv: arXiv:2303.14070. doi: 10.48550/arXiv.2303.14070.

Received June, 2024