Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 8, pp. 1733–1741
DOI : 10.47974/JSMS-1402

Designing a structured approach for consistency verification in AI systems for heartbeat disorder-related conversations

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

Large language models (LLMs) have shown mind-boggling question-answering capabilities. Most of their responses make one feel that they understand and remember context and are able to retrieve relevant content to generate almost accurate, human-like responses. While LLMs seem to mimic human intelligence, their responses can sometimes be inconsistent. Although LLMs are getting acceptance in healthcare for tasks like diagnosis support or patient communication, their inaccurate, inconsistent, or misleading information can be dangerous and can lead to harmful medical decisions if not properly validated by experts. Due to these limitations, LLM-powered full automation is still a dream. This paper presents a structured approach that can be used as basis to check LLMs level of "understanding" of medical knowledge and their "inferential capabilities" by evaluation of their responses to the word problems related to human heartbeat. The crafted word problems designed in the light of the proposed approach can be very helpful in identifying inherent weaknesses stemming from their fundamental nature. The study can be highly beneficial for healthcare professionals in determining the appropriate level of adoption of modern artificial intelligence (AI) technologies to identify heartbeat related disorders like Arrhythmia.

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

Keywords: *Large language models (LLMs), AI technologies, Healthcare, Structured approach, Heartbeat, Arrhythmia.*

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

1. Introduction

LLMs are one of the most incredible humans' achievements. Their advent is revolutionizing the landscape of intelligent automation. Industries are adopting chatbots that are built on the foundation of Large Language Models (LLMs). They leverage the advanced natural language processing capabilities of these foundation models and deliver human-like interactions. These chatbots have found their roots in number of industries and are used in domains like customer support, education, and healthcare.

Despite their wonderful capabilities, LLMs are not perfect and are seen to generate inaccurate, inconsistent and misleading information. Hence, blind reliance on these models can never be recommended. It is crucial to evaluate content generated by LLM. The discovered and reported weaknesses are helping AI professionals to produce better versions of LLMs.

Healthcare is a very sensitive area where human lives and their well-being are at stake. A very small mistake can lead to the misery of the patient and his / her family. One of the most sensitive areas in healthcare is heart care. Heartbeat related disorders like arrhythmias can lead to severe complications including stroke, heart failure, or sudden cardiac arrest. AI systems in this domain should be very consistent and correct in their responses. In this paper, we have used heartbeat related medical information as an example to build a structured approach to check consistency of any AI system. The proposed approach can provide guidelines using which a researcher can perform few systematic operations to discover inconsistency in the LLM response. Inconsistency in this paper refers to varying responses of AI system for queries that are same or slightly rephrased or are provide with some extra information, even though their answer remains same. For example, figure 1 will help in illustrating our definition.

The question that naturally arises is whether it would be a wise decision to adopt such technology in critical situations where human life is at stake. A structured approach that can help in evaluation of system's performance by providing series of standardized step to follow, can be very helpful to answer the discussed challenge. Using such approach, certain level of assurance can be provided to user that AI system will behave predictably when it will come across varying inputs, series of inputs and conditions.

Mr. Y was sitting with a heart rate of 50 bpm. He suddenly started running due to fear. What will likely happen to his heart rate immediately after stopping?

Options:

A. It will stay elevated between 70 bpm and 80 bpm for some time.

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

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

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

E. It will stay elevated between 110 bpm and 120 bpm for some time.

F. It will stay elevated between 120 bpm and 130 bpm for some time.

We found the answer to above query was changed when the question was repeated with same wording however, in different situations. In GPT-4o environment, when this question was asked first time, the answer was B. When the second time, the same query with different name was repeated in the same environment, the system answered with E. Third time, the query was altered with another name and two words i.e., "with haste" were added to word problem. The modified sentence was: "He suddenly started running due to fear with haste." Now the response of the system was F.

Figure 1

Example to show *inconsistent* behavior of LLMs using a word problem

In this paper, we present our work related to structured approach that can be used to check extent of understanding of medical knowledge of LLMs and their inferential capabilities by using human heartbeat related word problems. The study can assist decision makers especially in healthcare industry to determine the appropriate level of implementation of AI systems in sensitive areas for example to identify heartbeat related disorders like **Arrhythmia**.

The paper delivers information about related researches and similar works in its literature review section (section 2). The section is followed by section 3 discussing our methodology. The most important section 4 is the "proposed structure approach", where we present series of steps that can be helpful in evaluation an LLM for consistency. The final section (section 5) of paper i.e., conclusion summarizes the work and discusses future areas for improvement.

2. Literature Review

World is observing quick expansion of LLM ecosystem that is leading to an unprecedented surge in real-world AI applications. [1] reviews the evolution and impact of large language models (LLMs). However, [2] in

their key findings revealed limitations of LLMs in situations, such as small change in numerical values resulting in performance differences, sensitivity to question complexity level, and susceptibility to unrelated information. They used GSM8K benchmark for their research purpose. GSM8K was introduced to academic world by [3] and is a dataset of 8500 math word problems. [4] 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 founded on the researchers' observations that on some surprisingly trivial problems, models unexpectedly fail.

After such discussions as mentioned in [2], [4], the question that "can we trust such hallucinating models in healthcare industry?" demands clear answer. Even though, with all limitations and observed vulnerabilities, 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. For example [5] highlights the potential of ChatGPT in improving healthcare education, research, and practice. Similarly [6] 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 [7] 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.

[8] presents a comprehensive review of different methods to evaluate LLMs, with three key dimensions in focus: what to evaluate, where to evaluate, and how to evaluate. Their aim was to provide insights to assist in development of more proficient LLMs. Different frameworks are proposed with different purposes. Few frameworks discuss the mechanism to evaluate LLMs performance. They identified natural language understanding, reasoning, natural language generation, factuality and other areas to evaluate LLMs. Similarly [9] offers a comprehensive survey of key and crucial dimensions that should be consider when assessing LLM trustworthiness.[10] introduce LLM4Vuln, a unified evaluation framework to separate and assess LLMs' vulnerability reasoning capabilities and to examine improvements when combined with other

enhancements. Hence, it can be said that number of researchers are attempting to find ways that can help in identification of LLM vulnerabilities and this identification can result in solutions that can result in more consistent LLMs.

3. Methodology

In this section, we will briefly provide an overview of the methodology. Following steps were followed for the purpose of the formulation of structured approach in context of checking consistency of LLM based systems that may be working in medical domain.

1. Identification of the requirement of the structured approach (vulnerabilities of different existing LLMs were checked manually and simple approaches to discover such vulnerabilities were explored).
2. Survey of existing approaches (how researchers are discovering vulnerabilities?)
3. Determination of potential areas where the proposed approach can be applied
4. Selection of an area where proposed approach applicability can be demonstrated (In our case, it was health care domain in general and heartbeat related disorder area in particular)
5. Investigation of areas of medical knowledge related to heartbeat rates (for example: ability of LLMs to check variability of heartbeat rate in different states)
6. Exploration of dimensions that can be used to check LLM consistency (for example: time, space, heart related proverbs/idioms, statistical measures, etc.)
7. Formulation of word problems for identified areas of medical knowledge (Many types of patterns were made, but in this paper, we will present one pattern as an example for which few word problems were tested)
8. Querying different LLMs by giving word problems.
9. Evaluate the response of LLMs for consistency.
10. Analyze LLMs' responses and identify reasons why for certain problems, LLMs were unable to demonstrate consistent behavior

4. Proposed Structured Approach

This section will discuss the details of the proposed structured approach. Since the idea is to give confidence to health professionals about the consistency of the AI systems, the first step of this approach is to extract information form AI systems about their knowledge of medical domain. In our case, it will be about heartbeat disorders. Figure 2 describes part of medical information that was extracted after querying GPT-4o by asking about reasons for alternation of heartbeat.

The second step is to consult medical literature or domain experts to verify the provided information. This step is important because if the information provided to LLM has fault, then one should expect wrong answers from LLM. Even though, it is a rare possibility but one should be vigilant in performing this step too. In this step, identification of metrics

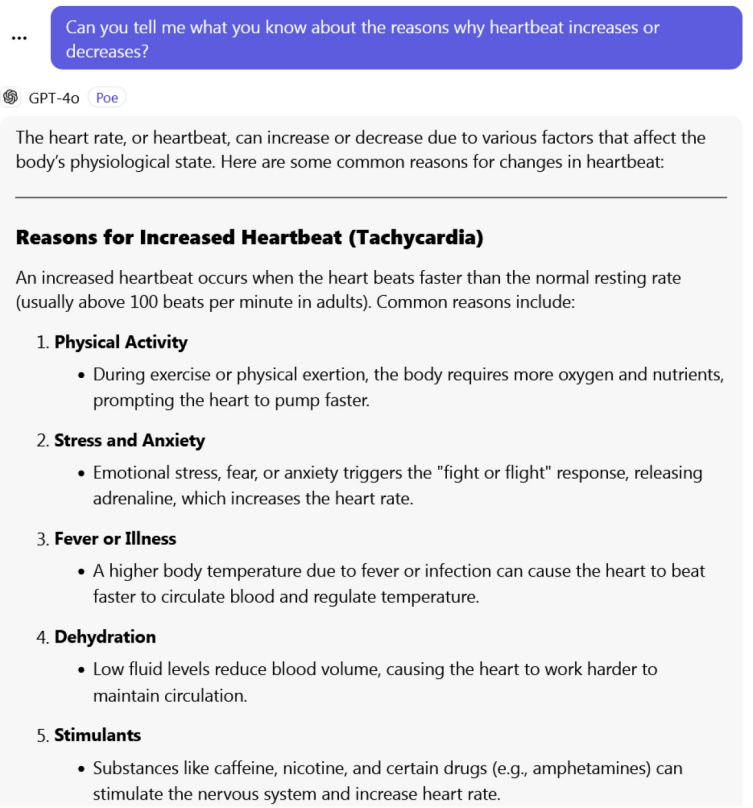

Figure 2

Extracting medical information from GPT-4o

and thresholds commonly used in heart-related problems should also be noted and LLMs should be asked about them. The output of this step should be the listing of medical topics that are needed to be evaluated in the next steps. Another way to check LLM knowledge is to ask it to do coding in which thresholds and ranges are used for different medical situations.

The next step is to align the extracted medical topics with statistical reasoning. Since statistics requires quantitative inputs therefore metrics and thresholds will be very useful input.

The next step is to create well-structured word problems combining statistical reasoning and heart-related medical knowledge to challenge LLMs. The patterns should be formulated that are able to generate multiple questions with multiple choices. For example, pattern in figure 3 can produce multiple questions.

It should be noted that since we are looking for consistency, it is not necessary that we have exact correct solution to the word problem. However, word problems with statistical measures will have exact solutions and for these questions, correctness of LMS responses can also be measured. The patterns should be formulated in such a way that different dimensions can be incorporated in the word problems that are derived from the patterns. For example, dimensions of time, space, English idioms and proverbs related to heart, adjectives and adverbs usage should be present in the patterns and combinations of one or more of them should be used to check LLMs vulnerabilities. Similarly in multiple choice questions, the choices positions should be changed to check impact on LLM or LLM based system response. For formulation of word problems that are intended to check the correctness of response, measures like mean, range, standard deviation should be properly placed in the pattern.

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

A. It will stay elevated between {y} bpm and {$y+10$} bpm for some time.

B. it will stay elevated between {z} bpm and {$z+10$} bpm for some time.

C. It will stay elevated between {i} bpm and {$i+10$} bpm for some time.

D. It will stay elevated between {j} bpm and {$j+10$} bpm for some time.

It should be noted that ranges of i, j, y and z are from 40 to 120 and i, j, y, z are not equal and should be 10 units away from each other.

Figure 3
An example of pattern that can generate multiple word problems

The next step is to check LLM performance based on the constructed synthetic dataset of word problems. It is expected that same question with different forms in different situations with different position of answer choices will enable exhaustive evaluation of AI system for consistency.

5. Conclusion

This paper presents pattern based structured approach that can be used to check the consistency of LLM in medical domain. Instead of using available general-purpose datasets, the approach can be helpful in formulation of domain-specific LLM constructed synthetic dataset(s) that is/are based on medical information provided by LLM itself. The future work will be to extend this work by implementing this approach to evaluate the consistency of different LLMs. The study can be highly beneficial for healthcare professionals in determining the consistency level of an AI system in general and particularly in context of heartbeat related disorders like Arrhythmia.

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] W. X. Zhao, X. Y. Li, Z. H. Wang, and A. B. Chen, "A survey of large language models," arXiv, Oct. 13 (2024), arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Y. Chang, Z. X. Li, M. P. Kim, and A. B. Zhao, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 39:1-39:45, Mar. (2024), doi: 10.1145/3641289.

[9] Y. Liu, S. M. Zhao, X. T. Zhang, and Q. W. Huang, "Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment," arXiv, Mar. 21 (2024), arXiv:2308.05374. doi: 10.48550/arXiv.2308.05374.

[10] Y. Sun, J. K. Wang, L. Z. Xu, and F. G. Yu, "LLM4Vuln: A unified evaluation framework for decoupling and enhancing LLMs' vulnerability reasoning," *arXiv*, Sep. 5 (2024), arXiv:2401.16185. doi: 10.48550/arXiv.2401.16185.

Received June, 2024