https://doi.org/10.62836/jcmea.v3i1.030106

Global Science Publishing

JCMEA

Journal of Computational Methods in Engineering Applications

https://ojs.sgsci.org/journals/jcmea

Article

AI-Driven Health Advice: Evaluating the Potential of Large Language Models as Health Assistants

Yanlin Liu [1,*] and Jiayi Wang [2]

[1] *Northeastern University, 4 N 2nd St Suite 150, San Jose, CA 95113, USA*

[2] *Wake Forest University, 1834 Wake Forest Rd, Winston-Salem, NC 27109, USA; wjyina@gmail.com*

Abstract: This study aims to evaluate whether the GPT model can be a health assistant by addressing health concerns from three aspects: providing preliminary guidance, clarifying information, and offering accessible recommendations. 31 questions in total were collected from multiple online health platforms, which included diverse health concerns across different age ranges and genders. A tailored system prompt was built to guide GPT model GPT-3.5-turbo generating responses. The evaluation metrics are designed based on 3 metrics: "Preliminary Guidance", "Clarifying Information", and "Accessibility and Convenience", which is used to evaluate responses with score method from 0 to 5. Lastly, the generated responses were evaluated using established metrics by an experienced medical doctor with over 20 years of experience in the fields of general and preventive care. The results indicate that LLMs demonstrated moderate performance in both the 'preliminary guidance' and 'clarifying information' aspects. Specifically, the mean score for 'preliminary guidance' was 3.65, implying that LLMs are capable of offering valuable insights when symptoms indicate the need for urgent or emergency care, as well as providing reassurance to patients for minor symptoms. In a similar manner, the mean score for 'clarifying information' was 3.87, demonstrating that LLMs effectively provide supplementary information to aid patients in making informed decisions. However, the mean score for 'accessibility and convenience' was notably lower at 2.65, highlighting a deficiency in LLMs' ability to offer advice customized to the specific needs of individual patients.

Keywords: large language models; LLMs; LLM; GPT models; GPT-3.5-turbo; artificial intelligence; healthcare; general health; health assistants; digital health

1. Introduction

Large Language Models (LLMs) such as GPT (Generative Pre-Trained Transformers) models are being used and impacted in a widely diverse field including healthcare [1]. They are able to understand complex tasks and generate human-readable text such as providing health advice and guidance [2–4]. The advent of AI Large Language Models (LLMs), spearheaded by OpenAI's Generative Pre-Trained Transformer (GPT), has sparked interest among people in academia and industry. The enhanced accessibility of logistics introduces a wide range of opportunities across sectors such as healthcare, finance, and retail, while progressing at an exponential rate. LLMs not only interpret and analyze data based on descriptive information and human directives, but they also possess the capability to process vast and complex datasets like textual documents and human dialogues,

Received: 15 October 2023; Accepted: 31 October 2023.

* Corresponding: Yanlin Liu (yanlinliu019@gmail.com)

generating coherent responses informed by their extensive training datasets [5]. In essence, individuals can interact with AI by providing natural language prompts, which guide LLMs in generating contextually appropriate responses [3,6]. Moreover, other methods can also facilitate this process [7–14].

Healthcare affects all individuals, with significant variations in the quality of care across geographic regions and socioeconomic groups [15], leading to disparities that contribute to uncertainty in life expectancy and overall quality of life, both of which are crucial to the well-being of the human race. While the full automation of LLMs without human oversight cannot yet be confidently applied across the medical field, in non-life-threatening situations, the role LLMs play in assisting professional healthcare providers warrants careful consideration [16]. The shortage of primary care providers, coupled with rising medical costs, has made patients hesitant to schedule timely appointments and to address vague or minor symptoms, particularly in the context of general and preventative care. Consequently, there is an urgent need for patients to access medical information generated by LLMs, similar to seeking a second opinion. This information can serve as guidance in their decision-making when uncertainties in medical knowledge arise during non-life-threatening situations, provided it adheres to certain quality standards.

Given their advanced language understanding capabilities, GPT models offer an effective approach to enhancing healthcare access through their role as virtual health assistants. LLMs can provide immediate health guidance and clarify potential diagnoses or treatment options to support patient decision-making processes [4]. This paper explores the potential of LLMs to aid patients in making informed health choices.

2. Data Collection

A total of 31 questions were gathered from multiple health-related open platforms, including Reddit, HealthUnlocked, Healthboards and Mayo Clinic Connect. During the data collection process, questions containing Personally Identifiable Information (PII) were filtered out to ensure that no sensitive data was passed to LLM [17]. In addition, the selected questions cover various general health topics and reflect a wide range of age groups to ensure data diversity.

3. Response Generation

The system prompt was developed to ensure that responses are precisely tailored to individual queries [3,18, 19]. Each response follows the structured sequence outlined by the system prompt to guarantee that the agent thoroughly understands and effectively addresses the patient's concerns [20]. This process includes providing preliminary guidance, clarifying information on potential conditions and treatments, and delivering pertinent recommendations. LangChain was utilized to enable the seamless integration of the system prompt with user queries, leveraging the GPT model, specifically gpt-3.5-turbo, to execute these tasks [21]. The final responses are stored and exported in an excel file, also each question and response are formatted as a Frequently Asked Questions (FAQ) document and stored in JavaScript Object Notation (JSON) format.

4. Evaluation

The generated responses were evaluated by a medical doctor based on three specific metrics: Preliminary Guidance, Clarifying Information, and Accessibility and Convenience. Each response was evaluated using a scale of 0 to 5 per metric, resulting in a possible total score between 0 and 15.

The preliminary guidance metric evaluated the degree to which the response offered insights into whether symptoms required urgent care, potentially preventing unnecessary visits or providing reassurance when appropriate. The clarifying information metric evaluated the response's capacity to deliver additional context regarding possible diagnosis or treatment options, thereby aiding the patient in making informed decisions [4]. The accessibility and convenience metric examined the extent to which the response provided practical and immediate advice [5], particularly in scenarios where access to healthcare providers was restricted due to appointment delays, scheduling conflicts, or high costs.

Each metric is equally weighted in determining the overall score, which is interpreted as follows: A score

between 0 and 5 indicates a poorly tailored response, lacking essential guidance, clarity, or accessibility. A score between 6 and 10 reflects a fairly tailored response, covering some elements but potentially lacking in completeness or depth. A score between 11 and 15 signifies a well-tailored response, comprehensively addressing all aspects of the chain of thought.

This scoring method offered a systematic method to assess the effectiveness with which the responses aligned with the intended objectives, ensuring a logical progression and comprehensive coverage of all required elements.

5. Results

Throughout the study, the GPT model consistently produced responses for all posed questions. A comprehensive record of responses corresponding to each question is provided in the [appendix 1]. The performance of these responses is depicted in the score distribution graph presented below.

The evaluation of the LLM's performance across the three metrics revealed that for Preliminary Guidance, the average score was 3.65 (±1.45), with 58% of the responses scoring above this average. For Clarifying Information, the mean score was 3.87 (±1.15), and 48% of the responses exceeded this average. For Accessibility and Convenience, the responses averaged 2.65 (±1.82), with 42% of them scoring higher than the mean. The Total Score had an average of 10.16 (±3.27), with 52% of the total scores surpassing this average. These findings indicate the proportion of LLM responses that outperformed their respective averages in each category.

The evaluation of the LLM's performance across the three metrics revealed distinct patterns. For Preliminary Guidance, the average score was 3.65 (±1.45), with 58% of the responses scoring above this mean. Clarifying Information had a mean score of 3.87 (±1.15), with 48% of responses exceeding this value. In terms of Accessibility and Convenience, the responses averaged 2.65 (±1.82), with 42% scoring higher than the mean. The overall total score showed an average of 10.16 (±3.27), and 52% of the responses surpassed this benchmark. These results demonstrate the proportion of LLM responses that performed above their respective mean scores within each category.

6. Discussion

The study aims to investigate whether LLMs can effectively provide essential health guidance as health assistants by leveraging the GPT model to address medical questions. A senior, experienced doctor was asked to evaluate the responses generated by the model. Overall, while the performance is well-tailored in terms of offering insights and clarifying information, the responses demonstrate limitations in delivering advice related to accessibility and convenience.

According to Figure 1 (Response Score Distribution) and Figure 2 (Response Performance Based on Evaluation Metrics), while it shows that more than half responses surpassed average score, the range from 2 to 14 highlights significant discrepancies in response quality. The average score 3.65 (±1.45) of metric "preliminary guidance" indicates that the GPT model performs moderate guidance by providing examples to elaborate differences between urgent symptoms and non-urgent symptoms. The 25th percentile value was 3, in other words, 75% of the responses have scores more than or equal to 3 which reflects adequacy of coverage. The average score 3.87 (±1.15) of metric "clarify information" demonstrates that the GPT model provides useful information on potential diagnosis or treatments. However, the lowest score is 1 with standard deviation 1.15 implies that there are some responses lacking detailed explanation. The metric "Accessibility and Convenience" has average score 2.65 (±1.82), which reflects that responses provide accessibility, but are not fully integrated into recommendations which should consider the key obstacle: high medical costs. The total score 10.16 (±3.27) demonstrates that the majority of responses were highly tailored to the question. Overall, the GPT model has provided excellent response, however, the responses need to be improved when considering the accessibility in the real world.

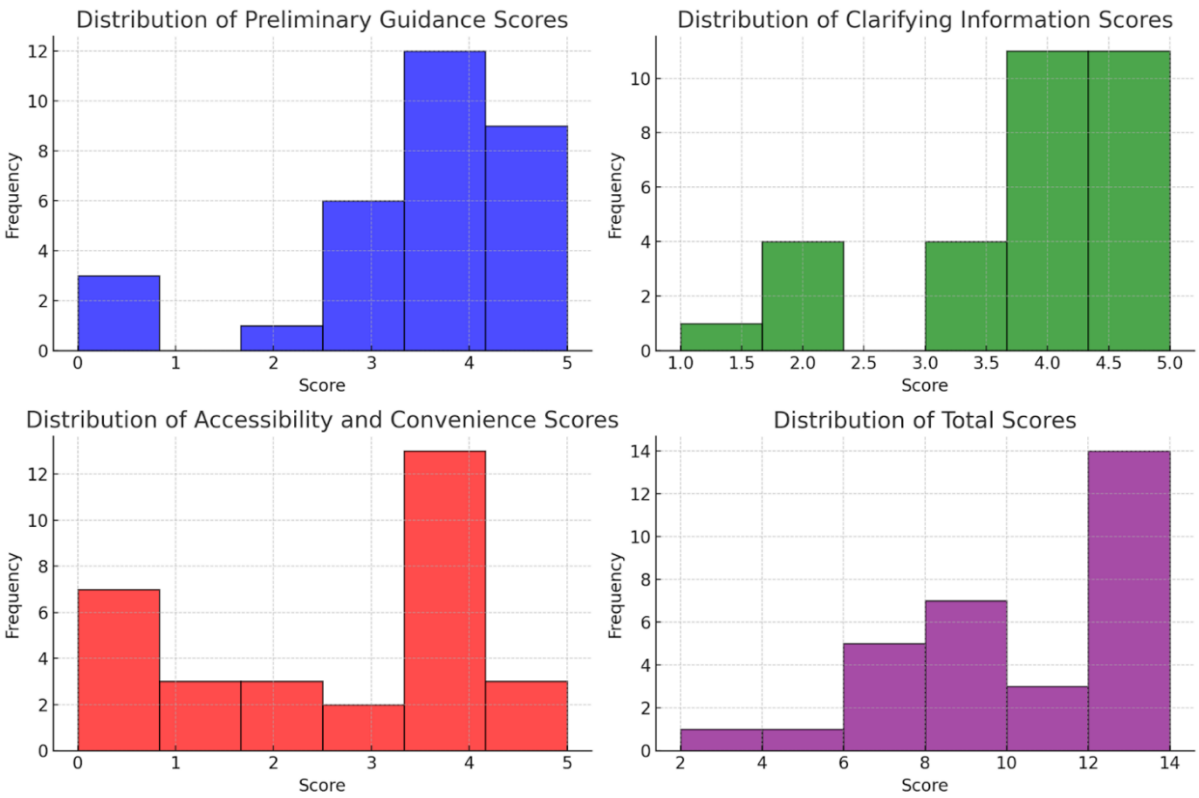

Figure 1. Response score distribution.

	Preliminary Guidance	Clarifying Information	Accessibility and Convenience	Total Score
count	31	31	31	31
mean	3.65	3.87	2.65	10.16
std	1.45	1.15	1.82	3.27
min	0.00	1.00	0.00	2.00
25%	3.00	3.00	1.00	8.00
50%	4.00	4.00	4.00	10.00
75%	5.00	5.00	4.00	13.00
max	5.00	5.00	5.00	14.00

Figure 2. Response performance based on evaluation metrics.

6.1. Limitation and Further Research

The first limitation of our study is the relatively small sample size, despite its coverage of multiple areas of general health and a wide range of age groups. Given the numerous sub-specializations within each health area, it is possible that not all are adequately represented, which may lead to underperformance in certain edge cases.

Additionally, the performance of the GPT model is inconsistent. The minimum score of 2 indicates responses lacking in essential guidance, clarity, and accessibility, while the maximum score of 14 reflects well-addressed aspects of the chain of thought, falling within the best response range (11 to 15).

The personalization of responses and evaluations can be improved. For example, the GPT model's ability to provide more personalized answers could be enhanced by incorporating patients' medical records as conversation history. Moreover, expanding evaluation metrics to assess how tailored the responses are to individual patient situations would provide a more comprehensive analysis.

Furthermore, conducting a comparative analysis offers a potential future direction. Evaluating multiple LLMs using the same questions, system prompts, and evaluation criteria would allow for a more robust comparison, particularly since variations in training datasets may lead to differing responses. In addition to comparing the GPT model with other LLMs, another valuable follow-up study would involve comparing LLM-generated responses with those of experienced human healthcare experts. Both LLMs and human experts could address the same set of health inquiries, with evaluations conducted by a senior doctor blinded to the origin of the responses.

7. Conclusion

This study demonstrates that the GPT model holds considerable potential as a health assistant by providing preliminary guidance and clarifying medical information, with more than half of its responses exceeding average evaluation scores. However, the model exhibits inconsistencies when delivering accessible recommendations, particularly concerning issues such as high medical costs. The variability in response quality underscores the need for enhanced personalization and more comprehensive evaluation metrics. Furthermore, the limited sample size and incomplete coverage across all medical sub-specializations indicate that further research is necessary to validate these findings. Overall, while the GPT model shows significant promise in supporting healthcare delivery, subsequent enhancements are crucial to fully realize its potential as a health assistant [16].

Funding

Not applicable.

Author Contributions

Y. L. and J. W. have made significant contributions, including the conception, design of the study, and the experimental framework. As the corresponding author, Y. L. was responsible for data collection, ensuring that non-sensitive information was passed to the GPT model. In addition, Y. L. implemented features to generate responses and ensured the accuracy and reliability of the development. J. W. took the lead in drafting the manuscript, coordinating revisions, and engaging experienced medical professionals to evaluate the quality of the generated responses. All authors provided critical insights during the interpretation of the results. The combined efforts of all authors were essential in successfully conducting this research and articulating its findings.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Sezgin E, Sirrianni J, Linwood S. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Medical Informatics* 2022; **10**(**2**): e32875.

2 Vinod V, Agrawal S, Gaurav VRP, *et al*. Multilingual Medical Question Answering and Information Retrieval for Rural Health Intelligence Access. *arXiv* 2021; arXiv:2106.01251.

3 Saxena S. *Medical Question Answering Using Instructional Prompts*; Arizona State University: Tempe, AZ, USA, 2021.

4 Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices. *NPJ Digital Medicine* 2018; **1**(**1**): 39.

5 Das A, Selek S, Warner AR, *et al*. Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues. In Proceedings of the 21st Workshop on Biomedical

Language Processing, Dublin, Ireland, 26 May 2022; pp. 285–297.

6 Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *arXiv* 2021; arXiv:2102.07350.

7 Chen Z, Fu C, Wu R, *et al*. LGFat-RGCN: Faster Attention with Heterogeneous RGCN for Medical ICD Coding Generation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 27 October 2023; pp. 5428–5435.

8 El Abbadi A, Dobbie G, Feng Z, *et al*. Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; pp. 331–343. https://doi.org/10.1007/978-3-031-35415-1_23.

9 Wang Y, Chen J, Wang M, *et al.* A Closer Look at Classifier in Adversarial Domain Generalization. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 27 October 2023; pp. 280–289. https://doi.org/10.1145/3581783.3611743.

10 Gu Y, Yan D, Yan S, *et al*. Price Forecast with High-Frequency Finance Data: An Autoregressive Recurrent Neural Network Model with Technical Indicators. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, 19 – 23 October 2020; pp. 2485 – 2492. https://doi.org/10.1145/3340531.3412738.

11 Gu Y, Chen K. GAN-Based Domain Inference Attack. *AAAI Conference on Artificial Intelligence* 2023; **37** (**12**): 14214–14222. https://doi.org/10.1609/aaai.v37i12.26663.

12 Gu Y, Sharma S, Chen K. Image Disguising for Scalable GPU-accelerated Confidential Deep Learning. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 3679–3681. Available online: https://dl.acm.org/doi/abs/10.1145/3576915.3624364 (accessed on 29 May 2024).

13 Du S, Chen Z, Wu H, *et al*. Image Recommendation Algorithm Combined with Deep Neural Network Designed for Social Networks. *Complexity* 2021; **2021**: 5196190. https://doi.org/10.1155/2021/5196190.

14 Wang Y, Chen Z, Fu C. Synergy Masks of Domain Attribute Model DaBERT: Emotional Tracking on Time-Varying Virtual Space Communication. *Sensors* 2022; **22**: 8450.

15 Shah TI, Clark AF, Seabrook JA, *et al*. Geographic Accessibility to Primary Care Providers: Comparing Rural and Urban Areas in Southwestern Ontario. *The Canadian Geographer/Le Géographe Canadien* 2020; **64**(**1**): 65–78.

16 Chintagunta B, Katariya N, Amatriain X, *et al*. Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization. *Machine Learning for Healthcare Conference* 2021; **149**: 354-372.

17 Libbi CA, Trienes J, Trieschnigg D, *et al*. Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records. *Future Internet* 2021; **13**: 136.

18 Valmeekam K, Olmo A, Sreedharan S, *et al*. Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). In Proceedings of the NeurIPS 2022 Workshop on Foundation Models for Decision Making, New Orleans, SL, USA, 5 October 2022.

19 Brown TB, Mann B, Ryder N, *et al*. Language Models are Few-Shot learners. *arXiv* 2020; arXiv: 2005. 14165.

20 Radford A, Wu J, Child R, *et al*. *Language Models Are Unsupervised Multitask Learners*; OpenAI Blog.: San Francisco, CA, USA, 2019.

21 Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research* 2020; **22**(**6**): e15154.