Annals of Biomedical Engineering
https://doi.org/10.1007/s10439-023-03327-6

BMES BIOMEDICAL ENGINEERING SOCIETY

LETTER TO THE EDITOR

Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation

Calvin Wang[1] · Joshua Ong[2] · Chara Wang[3] · Hannah Ong[4] · Rebekah Cheng[5] · Dennis Ong[6]

Received: 14 July 2023 / Accepted: 17 July 2023

Abstract

Advancements in artificial intelligence (AI) provide many helpful tools for healthcare, one of which includes AI chatbots that use natural language processing to create humanlike, conversational dialog. These chatbots have general cognitive skills and are able to engage with clinicians and patients to discuss patients' health conditions and what they may be at risk for. While chatbot engines have access to a wide range of medical texts and research papers, they currently provide high-level, generic responses and are limited in their ability to provide diagnostic guidance and clinical advice to patients on an individual level. The essay discusses the use of retrieval-augmented generation (RAG), which can be used to improve the specificity of user-entered prompts and thereby enhance the detail in AI chatbot responses. By embedding more recent clinical data and trusted medical sources, such as clinical guidelines, into the chatbot models, AI chatbots can provide more patient-specific guidance, faster diagnoses and treatment recommendations, and greater improvement of patient outcomes.

Keywords GPT · Patient · Clinical · Large language model

Introduction

Developments in artificial intelligence (AI) research continue to unearth key applications wherein AI can transform modern medicine. For example, prior studies have demonstrated AI's potential for analyzing medical images [1], coding medical notes [2], identifying high-risk patient groups [3], and summarizing clinical trials [4]. Another aspect where AI can revolutionize medicine is through AI chatbots, such as OpenAI's ChatGPT (OpenAI, San Francisco, CA) and Google's Language Model for Dialog Applications (LaMDA) (Google, Mountain View, CA).

While ChatGPT is not programmed for specific functions such as medical image interpretation and note generation, it holds general cognitive skills to engage in a conversation with chatbot users. Users can input "prompts" for the engine to respond to, which can take on many different forms such as a question (e.g., "How many layers exist within human skin?") or a directive (e.g., "Please summarize the following article into 3 main points.") [5].

AI chatbots are typically trained by publicly available sources on the Internet; for example, ChatGPT3.5 contains data up to September 2021. While ChatGPT has not been specifically trained for healthcare applications, it has the ability to access and reference a broad array of medical texts, research papers, health system websites, and clinical guidelines when prompted with a medical question [5]. In fact, ChatGPT is capable of exceeding the United States Medical Licensing Examination (USMLE) Step One passing score by approximately 20 points simply by leveraging its dataset [6].

However, while ChatGPT performs well in answering questions related to scientific knowledge, it is less useful in analyzing risk factors and characteristics of a patient to provide individualized diagnostic guidance and clinical

Associate Editor Stefan M. Duma oversaw the review of this article.

✉ Calvin Wang
cw1028@rutgers.edu

1 College of Medicine - Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ 08901, USA

2 Michigan Medicine, University of Michigan, Ann Arbor, MI, USA

3 Biotechnology High School, Freehold, NJ, USA

4 College of Medicine, The Ohio State University, Columbus, OH, USA

5 Department of Physical Therapy, Virginia Commonwealth University, Richmond, VA, USA

6 Amazon Web Services, Amazon, Seattle, WA, USA

Published online: 02 August 2023

advice. In these circumstances, ChatGPT typically responds to prompts with less specificity [7].

For example, Fig. 1 illustrates ChatGPT's response to a hypothetical patient with respiratory issues, wherein the application suggests potential diseases and conditions that may be affecting the patient. Although the engine returns a comprehensive answer, it is generic and does not drive toward understanding more specific information about the patient (e.g., biomarkers) to provide a patient-specific recommendation or diagnosis.

Enabling this functionality within AI chatbots would increase the scale and speed at which patients can be diagnosed. This would be greatly helpful for clinicians across all specialties, but specifically within oncology where every delayed month in treatment can raise the risk of death by 6–13% [8]. Thus, ChatGPT would create an immense value-add by providing efficient and accurate diagnostic advice on a one-to-one basis for patients based on clinical characteristics.

One method through which ChatGPT can improve its response specificity is through retrieval-augmented generation (RAG), as shown in Fig. 2A. To allow RAG to occur, knowledge sources (e.g., document repositories, APIs) outside the limits of the model must first be added to

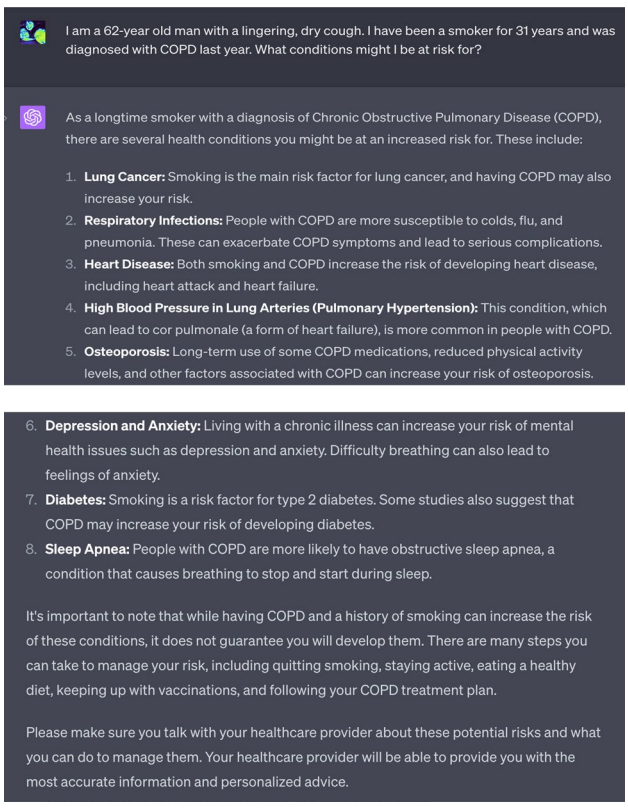

Fig. 1 Current ChatGPT4 conversation providing clinical advice for a hypothetical patient

expand the model's database. This addition of outside data is completed through embedding, or the process by which text is given numerical representation to be understood within a vector space. As a result, user prompts and queries (step 1) can draw upon these new knowledge sources (steps 2 and 3) in order to acquire "enhanced context" which can be used to augment the prompt and make it more specific (step 4). The augmented prompt may thereby generate a more detailed response from the large language model (step 5) [9].

As an additional example, Fig. 2B demonstrates how the RAG process works with the hypothetical question from Fig. 1. In the example, the original question (step 1) is asked as a prompt by the user and subsequently entered as a query in the system (step 2). The query is sourced through all available knowledge resources embedded within the model via RAG (step 3). In this case, additional context suggests that the patient's baseline condition resembles common risk factors for lung cancer. This context can then be used to augment the user's original prompt to be more focused on lung cancer (step 4), and thus the large language model can provide a more specific answer regarding biomarkers, procedures, and other diagnostic processes used to evaluate the disease. This level of retrieval-augmented generation, paired with prompt engineering, holds the potential to offer medical insights to clinicians and patients at a quick rate to guide clinical decision-making.

Figure 3 provides a bare, preliminary glimpse on what RAG-assisted ChatGPT responses could look like. Based on the same original question in Fig. 1, the prompt inputted below includes the enhanced context in step 3 of Fig. 2B. As displayed, this additional context allows ChatGPT to generate a response that is more detailed than that of Fig. 1. Here, ChatGPT lists additional symptoms, risk factors, and diagnostic tests that can be used to confirm a case of lung cancer. Via RAG, future ChatGPT engines would automatically retrieve this enhanced context to augment the original prompt and generate a more specific response. However, one can notice that the current response in Fig. 3 still does not contain advice to assist with diagnoses and treatments for patients on an individualized level, and future use of RAG with patient-specific data may help bridge this gap.

Therefore, the many advancements within healthcare from artificial intelligence can be brought one step forward through the embedding of more recent, detailed clinical data in chatbots to assist with retrieval-augmented generation. Clinical guidelines released after ChatGPT's cutoff of September 2021 and trusted medical sources (e.g., NCCN, UptoDate) can be added to models to augment user prompts and improve the specificity of ChatGPT's responses, in order to provide more patient-specific guidance to improve diagnostic and treatment efficiency.

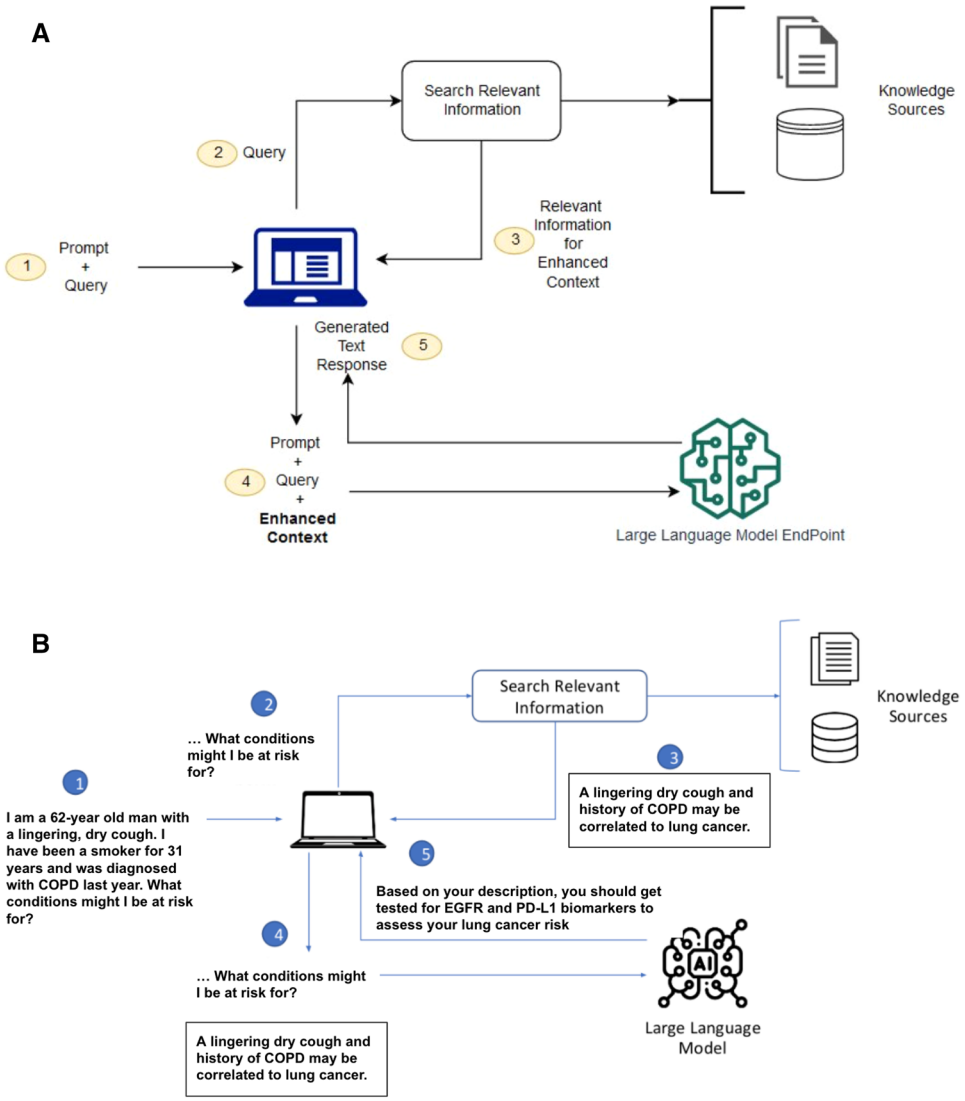

Fig. 2 **A** Illustration of retrieval-augmented generation (RAG) in a large language model, wherein knowledge sources are used to augment a prompt in order to generate a more accurate response from the model [10]. **B** Example of retrieval-augmented generation using the prompt from Fig. 1

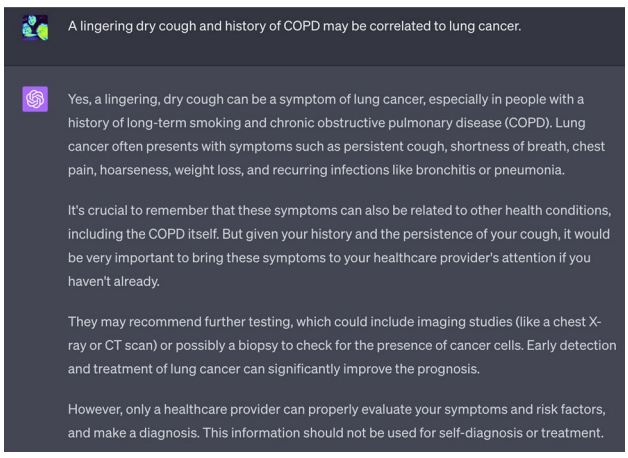

Fig. 3 Extension of ChatGPT conversation in Fig. 1, wherein the enhanced context from Fig. 2B is added to increase ChatGPT response specificity.

Funding Not applicable.

Declarations

Conflict of interest This submission does not contain any conflicts of interest or competing interests with prior submitted papers.

References

1. Ker, J., L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. *IEEE Access*. 6:9375–9389, 2018.
2. Milosevic, N., and W. Thielemann. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *J. Web Semant.* 2022. https://doi.org/10.1016/j.websem.2022.100756.

3. Beaulieu-Jones, B. K., W. Yuan, G. A. Brat, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit. Med.* 4:62–62, 2021.
4. Waisberg, E., J. Ong, M. Masalkhi, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J. Med. Sci.* 2023. https://doi.org/10.1007/s11845-023-03377-8.
5. Lee, P., S. Bubeck, and J. Petro. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* 388(13):1233–1239, 2023. https://doi.org/10.1056/nejmsr2214184.
6. Nori, H., N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of GPT-4 on medical challenge problems. arXiv: https://arxiv.org/abs/2303.13375, 2023.
7. Hanna, T. P., W. D. King, S. Thibodeau, M. Jalink, G. A. Paulin, E. Harvey-Jones, D. E. O'Sullivan, C. M. Booth, R. Sullivan, and A. Aggarwal. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ*. 2020. https://doi.org/10.1136/bmj.m4087.
8. Homolak, J. Opportunities and risks of chatgpt in medicine, Science, and Academic Publishing: a modern promethean dilemma. *Croat. Med. J.* 64(1):1–3, 2023. https://doi.org/10.3325/cmj.2023.64.1.
9. Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* 33:9459–9474, 2021.
10. Mishra, A. Machine Learning in the AWS Cloud: Add Intellegence to Applications with Amazon Sagemaker and Amazon Rekognition. Amazon, 2019. https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html.