Clinical eHealth 7 (2024) 147–152

Contents lists available at ScienceDirect

Clinical eHealth

journal homepage: ww.keaipublishing.com/CEH

Conversational AI with large language models to increase the uptake of clinical guidance

Gloria Macia [a,*], Alison Liddell [b], Vincent Doyle [c]

[a] Lead Data Engineer, London School of Economics and Political Science (LSE), UK
[b] Deputy Director of Digital Information & Technology Team, The National Institute for Health and Care Excellence (NICE), UK
[c] Product Development, Digital Team, The National Institute for Health and Care Excellence (NICE), UK

ARTICLE INFO

Article history:
Received 25 February 2024
Revised 8 August 2024
Accepted 10 December 2024
Available online 13 December 2024

Keywords:
Large Language Models
ChatGPT
Conversational AI
Clinical guidance
Retrieval-Augmented Generation (RAG)
National Institute for Health and Care Excellence (NICE)

ABSTRACT

The rise of large language models (LLMs) and conversational applications, like ChatGPT, prompts Health Technology Assessment (HTA) bodies, such as NICE, to rethink how healthcare professionals access clinical guidance. Integrating LLMs into systems like Retrieval-Augmented Generation (RAG) offers potential solutions to current LLMs' problems, like the generation of false or misleading information. The objective of this paper is to design and debate the value of an AI-driven system, similar to ChatGPT, to enhance the uptake of clinical guidance within the National Health Service (NHS) of the UK. Conversational interfaces, powered by LLMs, offer healthcare practitioners clear benefits over traditional ways of getting clinical guidance, such as easy navigation through long documents, blending information from various trusted sources, or expediting evidence-based decisions in situ. But, putting these interfaces into practice brings new challenges for HTA bodies, like assuring quality, addressing data privacy concerns, navigating existing resource constraints, or preparing the organization for innovative practices. Rigorous empirical evaluations are necessary to validate their effectiveness in increasing the uptake of clinical guidance among healthcare practitioners. A feasible evaluation strategy is elucidated in this research while its implementation remains as future work.

1. Introduction

Implementation research is the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and hence, to improve the quality and effectiveness of health services and care[1]. By summarizing the best available evidence for a particular clinical condition and providing clear recommendations for practice, clinical guidance can help healthcare professionals make informed decisions about the care they provide to their patients.

Despite the availability and accessibility of clinical guidance, ensuring that healthcare professionals consistently utilize these recommendations in their daily practice remains a significant challenge. The National Institute for Health and Care Excellence (NICE) plays a crucial role in developing and disseminating clinical guidance for the NHS and the wider health and care system. NICE defines its guidance as "Evidence-based recommendations for the health and social care sector, developed by independent committees, including professionals and lay members, and consulted on by stakeholders". These recommendations are widely used by healthcare practitioners in the UK and often adapted for use abroad[2].

However, traditional methods of disseminating clinical guidance, such as websites and downloadable PDFs, may not be sufficient to ensure that healthcare professionals can easily access and apply this information in a timely and effective manner. This challenge is particularly acute in fast-paced clinical environments where quick access to relevant information is crucial.

Artificial intelligence (AI), particularly large language models (LLMs), have opened new possibilities for enhancing the way healthcare professionals engage with clinical guidance[3]. Conversational AI, driven by these models, has the potential to provide instant, interactive, and context-aware access to clinical recommendations[4]. By offering a more intuitive and accessible medium for engaging with clinical guidance, conversational AI could

Abbreviations: LLM, Large Language Model; RAG, Retrieval Augmented Generation; NICE, National Institute for Health and Care Excellence; Gen AI, Generative Artificial Intelligence; HTA, Health Technology Assessment; NHS, National Health System.

* Corresponding author.
E-mail address: g.macia-munoz@lse.ac.uk (G. Macia).

https://doi.org/10.1016/j.ceh.2024.12.001

significantly improve the uptake and implementation of these guidelines in clinical practice.

The importance of this research lies in exploring how conversational AI can be leveraged to overcome current barriers in the dissemination and utilization of clinical guidance in the UK. Enhancing the engagement of healthcare professionals with clinical guidelines through AI could lead to better adherence to best practices, ultimately improving patient outcomes and healthcare quality.

2. Objective

The objective of this paper is to design and debate the value of an AI-driven system, similar to ChatGPT, to enhance the uptake of clinical guidance within the National Health Service (NHS) and the wider health and care system. This addresses one of the four pillars of NICE's five-year strategy (2021–2026)[5]: increasing the uptake of medical guidance in the UK to improve healthcare quality and effectiveness. As part of NICE's efforts to disseminate evidence-based recommendations, it is imperative for the institute to assess whether leveraging large language models (LLMs) can reimagine and improve the engagement of healthcare professionals with clinical guidance.

This research consists of the design and implementation of the AI-driven system. It also involves a debate among two senior executives from NICE and an AI expert from the London School of Economics and Political Science (LSE) on the potential benefits and drawbacks of such a system.

Evaluating the impact of this AI-driven system is a complex and critical task, given the NHS's vast scale and significance. In 2023/24, there were an estimated 600 million patient contacts with GP, community, hospital, NHS 111, and ambulance services, equating to 1.7 million interactions with patients daily[6]. The NHS also employs 1.5 million people, making it the UK's largest employer and one of the largest globally by headcount[7]. These facts underscore the immense potential benefits of improving clinical guidance uptake, such as better patient outcomes and increased operational efficiency. Nevertheless, they also highlight the substantial risks if the system were to provide inadequate information.

Therefore as an additional research objective we present the design of the AI-system evaluation. The evaluation of the system itself is, however, not included in this manuscript. This acknowledged limitation is a common practice in implementation research, where national-scale, multi-year interventions are divided into separate phases for the intervention design, implementation, and evaluation.

3. Materials and methods

3.1. Large language models as building Blocks

Large Language Models (LLMs) are cutting-edge AI technologies with the potential to revolutionize the medical field. Under the hood, LLMs are mathematical functions loosely inspired by the structure of the human brain. LLMs range from hundreds of millions to trillions of parameters which are learned from extensive text data during the training process. These advanced models offer powerful natural language capabilities such as text generation, translation, and question-answering with many possible applications to the medical field [8,9].

ChatGPT is a conversational application that utilizes a large language model (LLM) like GPT-4 as its foundation [10]. Consider the following analogy: If ChatGPT were a car, GPT-4 would be its engine. As even better LLMs emerge, the team behind ChatGPT can seamlessly swap out the engine without interrupting the users' experience as conversations will remain stored in the same database (the trunk), and users will continue to interact with the LLM via the same graphical user interface (the chassis). Another way to grasp GPT-4 is to think of it as a building block from a construction set. By combining the LLM with other components in an architecture, an endless amount of applications can be built to achieve different purposes.

3.2. The retrieval-augmented generation architecture

One of the most popular recent architectures is the Retrieval-Augmented Generation (RAG) architecture[11]. ChatGPT can answer a wide range of questions, but this versatility comes with known limitations, including occasional hallucination (the generation of incorrect or fictional information), dependence on static training data (as it lacks access to real-time information), and controllability issues (a model trained with public content from the entire internet may lead to biased or inappropriate content generation), raising concerns about ethical use in clinical practice. Hence, ChatGPT alone would not be a wise implementation strategy to increase the uptake of clinical guidance.

By contrast, the RAG architecture combines the text generation strengths of LLMs with the ability to retrieve information exclusively from a trusted knowledge base, such as NICE guidance. RAG works in a two-step process: retrieval and generation.

In the retrieval step, the user's query is converted into a numerical vector representation, known as an embedding, and compared to embeddings of text chunks extracted from the knowledge base. The most relevant chunks are then retrieved, providing the LLM with a deeper understanding of the context[11].

In the generation step, the retrieved context is augmented to the original query, forming a more comprehensive prompt for the LLM[11]. The LLM then generates a response based on the enriched prompt, ensuring it is tailored to the specific context and provides accurate information.

While this approach narrows the range of questions a chatbot can answer, it enhances factual accuracy, provides up-to-date information without retraining the LLM (minimizing costs), and improves controllability, rendering it a more suitable tool for clinical question-answering. In the following section, we will debate the advantages and disadvantages of utilizing LLMs as key components of RAG architectures as an implementation strategy to increase the update of medical guidance. A detailed technical diagram of the architecture is included in the results section.

4. Results

4.1. Proposed architecture of a conversational AI system to increase the uptake of NICE guidances in the UK

The architecture is a multi-tenant, microservices-based system orchestrated with Amazon Elastic Kubernetes Service (Amazon EKS), designed for robust, scalable AI-driven interactions[13]. It features a load balancer for distributing network traffic and an ingress controller for managing external access. Each tenant has its own proxy for authentication, ensuring secure and isolated access. Tenant-specific chatbots and vector databases handle the unique data and interaction needs of each tenant, managed through the LangChain framework[14] (Fig. 1).

The system integrates AI foundation models to provide intelligent and contextually relevant responses. Additionally, session and chat history data are stored using Amazon DynamoDB (a fully managed NoSQL database), ensuring persistent and retrievable interactions. The process flow begins with a user request routed through the load balancer and ingress controller to the identity

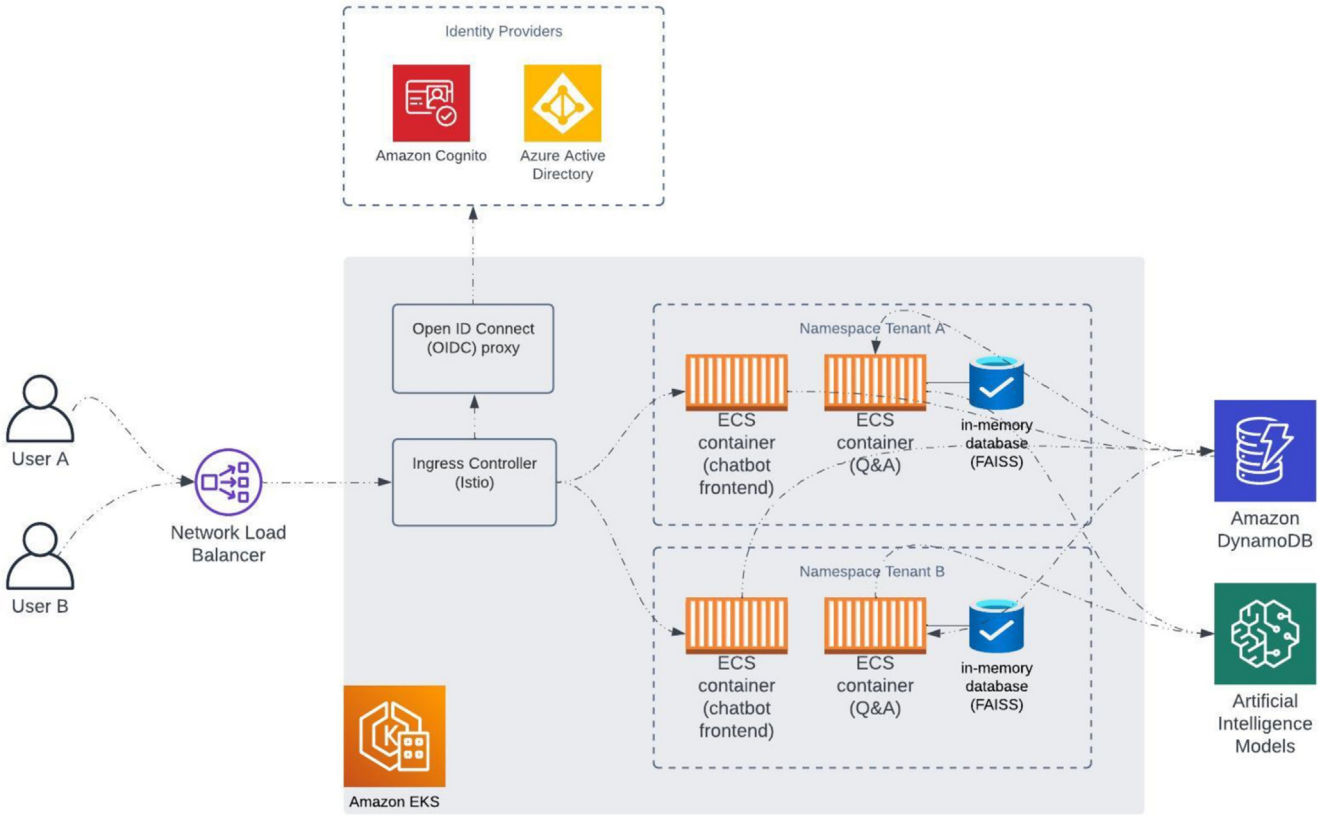

Fig. 1. Architecture Diagram of RAG-enabled AI Conversational System. A scalable, multi-tenant RAG-enabled conversational AI system orchestrated with Amazon Elastic Kubernetes Service (Amazon EKS). Key components include a network load balancer, an ingress controller for traffic management (istio) and tenant-specific proxies for secure authentication. Authenticated requests are forwarded to tenant-specific namespaces that consist of two component microservices built as Docker container images: a frontend chat container interface that interacts with the client and a question and answering (Q&A) container with an in-memory vector database (Facebook AI Similarity Search, FAISS) for efficient retrieval. OpenAI models are used to convert textual data to vector embeddings, and to generate the response to user queries. Session data is stored in Amazon DynamoDB for persistence. This architecture ensures robust, secure, and scalable AI interactions across multiple tenants.

provider for authentication. Upon successful authentication, the request is directed to the tenant-specific namespace. The session ID and chat history are managed via DynamoDB tables, maintaining data integrity and continuity. The chatbot then interacts with foundation models to generate accurate responses, which are sent back to the user.

This architecture ensures scalability, secure multi-tenancy, and flexibility, allowing individual components to be managed and updated independently. It supports isolated tenant environments, making it well-suited for complex applications requiring robust, scalable, and secure AI-driven interactions across multiple users.

Not depicted in this architecture diagram is the data ingestion pipeline that transforms NICE guidance into a format suitable for use within the system. This pipeline leverages LlamaIndex[15] to efficiently extract clinical tables, charts, and images from the NICE guidance documents (PDFs). While the footers of pre-extracted graphic elements are directly indexed in the vector database, the actual images are first sent to a multimodal LLM (GPT-4o[16(p4)]) to generate a textual description. This model generates responses based on the vectorized user question, interaction history, and the full context provided by the images themselves, rather than relying solely on the vectorized metadata or footers. This ensures that the responses are comprehensive and accurately reflect the detailed content of the images.

Embeddings computed with *text-embedding-3-large* [17] model are stored in a in-memory vector database (Facebook AI Similarity Search, FAISS[18]), facilitating efficient retrieval based on cosine similarity metric. The system typically retrieves 5 to 10 relevant text fragments for the LLM, adjusting dynamically based on query complexity to balance inference speed and response accuracy.

Given that the full AI system has not yet undergone evaluation and cannot be publicly released, the authors have provided a much simplified version for readers to explore. This version, built using Chatbase (a commercial custom AI chatbot builder) is accessible through a web portal with the sole objective to enhance the comprehensibility of the other sections. It exemplifies a conversational AI chatbot utilizing the RAG architecture, the GPT-3.5 model, and a single NICE guidance, the "Caesarean birth" guidance (NG192). It is not representative of the actual system performance, but it is useful for providing the reader with an intuitive idea.

4.2. Evaluation design of the conversational AI system to increase the uptake of NICE guidances in the UK

This section details the framework for evaluating the AI system. Notably, the actual assessment of the system's performance is not included in this manuscript. This recognized limitation aligns with standard practices in implementation research, where extensive, multi-year projects are divided into phases: design, implementation, and evaluation. The necessity for iterative development, ethical approvals, funding limitations, and logistical challenges inherent in large-scale healthcare interventions drive this phased approach. Future research should be conducted to empirically evaluate the AI system's impact on clinical guidance adherence and healthcare outcomes, aiming for a comprehensive assessment of its effectiveness and scalability within the NHS framework.

Evaluating the influence of this AI-driven system is a complex and essential task due to the vast scale and significance of the NHS. In 2023/24, the NHS experienced an estimated 600 million patient interactions across GP, community, hospital, NHS 111, and ambulance services, equating to approximately 1.7 million daily patient contacts[6]. These figures underscore the immense potential benefits of enhancing clinical guidance uptake, such as improved patient outcomes and increased operational efficiency. However, they also highlight the substantial risks involved if the system provides inadequate information. Previous research underscores the significance of detection mechanisms to filter out inaccuracies and hallucinations[19].

To this aim we propose the LLM-as-a-judge framework[20], which has demonstrated potential in automating and scaling the evaluation of language models[21,22]. The primary objective of this novel evaluation framework is to achieve a high correlation between the LLM's ratings and human ratings, thus validating the LLM-as-a-judge framework as an effective tool for assessing the performance of the RAG system.

The initial step involves the creation of a human evaluation dataset. This dataset will consist of a representative sample of question–answer pairs derived from the NICE guidelines, each evaluated by clinical experts. The pairs will be rated based on criteria such as relevance, accuracy, and comprehensiveness, with these ratings serving as a benchmark for subsequent evaluations. The agreement between human raters will be assessed using statistical measures to establish the reliability of these ratings as ground truth.

To implement the LLM-as-a-judge framework, the architecture uses a second LLM to act as an evaluator of the system's responses. The evaluation prompt for the LLM includes a detailed task description, a clear definition of the rating scale, and several examples of high-quality responses to guide the evaluation process.

The LLM will rate the system's answers on a scale, considering factors such as adherence to NICE guidelines, clarity, and clinical relevance. This approach leverages the LLM's ability to provide consistent and scalable evaluations, which would be time-consuming and costly if performed solely by human judges.

To ensure the reliability of the LLM judge, its ratings will be compared with those provided by human experts. This comparison will involve statistical measures such as Pearson correlation to determine the level of agreement. Iterative refinements will be made to the LLM's prompt and rating criteria to enhance alignment with human judgments.

This approach offers several key advantages, including scalability, cost-effectiveness, and consistency. The LLM-as-a-judge framework enables large-scale evaluations that are impractical for human judges alone, reducing both time and cost. It ensures consistent and unbiased ratings through predefined criteria. Additionally, the flexibility to swap LLM models, allows the system to easily integrate advancements in AI technology, enhancing its adaptability.

4.3. Pros and cons of the implementation of a conversational AI system to increase the uptake of NICE guidances in the UK

4.3.1. Efficient navigation through lengthy documents

Understanding how users need to receive information for the most efficient impact on their information uptake is crucial to guidance-issuing bodies. While many healthcare practitioners directly access the NICE website and navigate to the appropriate guidance page, the challenge lies in efficiently locating specific information within a lengthy document. For instance, the NICE guidance "Early and locally advanced breast cancer: diagnosis and management" [NG101] spans over 80 pages[23]. A new way of interfacing would allow guidance-issuing bodies to be much more targeted which could translate to a better user experience and ultimately better outcomes for patients.

4.3.2. Integration of information from multiple trusted sources

The RAG architectures present a unique opportunity to integrate information from multiple trusted sources seamlessly. This could include combining information from several NICE guidance documents (especially relevant for patients with comorbidities and/or in special situations like pregnancy or breastfeeding) and other trusted sources such as guidance from medical societies or other health technology assessment bodies. Moreover, since LLMs are able to translate text of clinical guidance documents to other languages, the language of the source documents does not necessarily have to be English or shared. Similarly, even if the original guidances are in English, the user could interact with them in another language, fostering cultural diversity and inclusion in the healthcare workforce [24]. It is important to note that the accuracy of translations into languages other than English depends on the capabilities of the underlying LLM. Consequently, this aspect should be thoroughly explored in the system's evaluation. As NICE guidance documents are exclusively in English, this study has not explored translation capabilities.

4.3.3. Up-to-date Insights for practitioners

Clinical guidance documents are continuously updated to reflect the newly generated medical evidence. This constant evolution poses a challenge for healthcare practitioners to stay up to date with the latest knowledge. Notably, RAG-enabled LLMs present a distinctive advantage in this scenario, as they do not necessitate retraining to seamlessly incorporate the most recent version of clinical guidance. Returning to the previous car analogy, the reviewed clinical guidance will be swapped out in the database (the trunk) without interfering with the engine (the LLM) and consequently impacting the driver's experience. This characteristic empowers practitioners by providing easy access to up-to-date knowledge directly from any device, streamlining the process of staying informed in the fast-paced landscape of medical advancements.

4.3.4. Decisiveness where it matters

In busy emergency rooms and critical care situations, healthcare professionals often have to make quick decisions that can directly affect patients. While being able to efficiently find information in lengthy clinical guidance documents is helpful during a regular 10-minute visit to a general practitioner, it becomes crucial in other medical scenarios, where it could be life-saving. This is especially true for A&E doctors who usually only have their phones on hand. Though it has not been evaluated as part of this research, we believe RAG-enabled LLMs could change the way healthcare practitioners interact with clinical guidance, making it easy to access from any device and enabling quick evidence-based decisions right where they are needed.

4.3.5. Assuring quality in the implementation of novel technologies

The task at hand for any guidance-issuing body like NICE is to guarantee that any system presenting its guidance is adequately trained and undergoes rigorous quality assurance. From experiments, the organization has learned that while these systems are impressive they are not yet perfect[12]. These technologies are still very new and rapidly evolving. Even if the benefits outweigh the risks, NICE, as a provider of crucial information to healthcare providers, must maintain confidence in the outcomes generated by the system to prevent unintended consequences. Prioritizing risk management and maintaining high-quality standards are essential to uphold the integrity of the provided guidance.

4.3.6. Emerging data privacy concerns

The utilization of a system by clinical practitioners to address medical queries inevitably introduces apprehensions regarding data privacy and security. Sharing sensitive patient information with external entities, such as OpenAI, increases the risk of data breaches, unauthorized access, and potential misuse of personal data [25]. Fortunately, there are strategies for organizations to leverage LLMs without violating data privacy laws. One approach involves hosting the LLM on internal servers, ensuring the highest security by eliminating data transmission to third parties. Alternatively, some companies offer commercial LLMs that prioritize data privacy. This allows organizations to set up agreements through contracts to protect sensitive information.

4.3.7. Balancing Ambition with due Diligence

As a public body and a relatively small organization, NICE like other HTA bodies is compelled to find the balance between where to invest most appropriately for the value it offers. Despite harboring significant strategic ambitions, the organization recognizes the need for thorough investigation and consideration of the available investment scale. Organizations must carefully assess where to allocate resources to derive maximum value, ensuring a measured and iterative approach to implementation, starting small and progressively advancing based on lessons learned. This cautious strategy is essential to make informed decisions on how to leverage LLMs effectively within the limitations of the organization's resources.

4.3.8. Navigating organization readiness challenges

Historically NICE's advanced analytics initiatives have primarily focused on automation and machine learning along its evidence synthesis and guidance development pipeline, particularly in tasks such as deduplicating studies. Building on these experiences and the recent surge in Generative Artificial Intelligence (Gen AI), NICE recognizes the importance of integrating AI into its overall organizational strategy. Taking a collaborative approach, the public body is consistently learning, pinpointing use cases suitable for this technology, and engaging with key players in the tech industry interested in collaboration. NICE is currently undergoing a digital transformation journey. While it has some excellent advanced analytics people in-house, the HTA body is considering how to organize to empower them to drive innovation across the organization.

4.1.8. Dealing with the costs and liability

NICE's current liability framework and budget were not planned to accommodate the extensive demands of deploying and maintaining a national AI system. If a conversational AI system leveraging NICE guidelines were to be rolled out to increase the uptake of clinical guidance in the UK, a critical question to address is which entity would be liable for its recommendations to clinical practitioners. As previously mentioned, in 2023/24, there were an estimated 600 million NHS patient contacts, equating to 1.7 million interactions with patients daily[6]. The cost of running such an AI system in production would undoubtedly exceed NICE's budgetary constraints.

Additionally, the costs associated with evaluating the system prior to deployment (e.g. evaluation costs), as well as ongoing improvements and performance monitoring once in production, are significant considerations that must be addressed and which pose, in our opinion, the main challenges for the viability of the project. Establishing collaborative efforts between governmental agencies, healthcare providers, and possibly private sector partners may be necessary to share the financial and legal responsibilities associated with the system's deployment and operation.

5. Discussion and Conclusions

This publication proposes the application of LLMs in conversational interfaces as a new strategy to increase the uptake of clinical guidance. We presented the architecture of a RAG-enabled conversational AI system, the design of its evaluation and the pros and cons that part of the NICE leadership team foresees in rolling it out for wider NHS use.

We believe that embracing LLMs in architectures like RAG can potentially mitigate current LLM drawbacks, such as generating false or misleading information due to biases, outdated training data, or misinformation embedded in public sources. Furthermore, conversational interfaces can offer substantial advantages over current methods for accessing clinical guidance. These advantages encompass the ability to navigate lengthy documents efficiently, integrate information from various trusted sources, stay informed about the latest medical knowledge, and facilitate rapid evidence-based decisions in situ. However, the implementation of such interfaces introduces new challenges for HTA bodies (managing risks and ensuring quality, addressing data privacy concerns, coping with existing resource constraints, and preparing the organization to drive innovation). An acknowledged limitation of this research is that while the design of the evaluation strategy was elucidated, we lacked the funding for further empirical research to evaluate its effectiveness in real-world healthcare settings.

The authors believe that though there are risks and challenges in the application of LLMs in conversational interfaces to access clinical guidance, potential benefits could outweigh them. Past studies have already concluded that LLMs like ChatGPT struggle to provide accurate responses to medical questions [26]. We argue that LLMs in an RAG architecture may solve many of these known limitations of LLMs. As a result, RAG-enabled LLMs conversational AI systems could facilitate the systematic integration of research findings and evidence-based practices into everyday clinical practice, increasing the uptake of clinical guidance among healthcare practitioners, and ultimately enhancing the quality and effectiveness of healthcare services.

CRediT authorship contribution statement

Gloria Macia: Writing – review & editing, Writing – original draft, Software, Conceptualization. **Alison Liddell:** Writing – review & editing. **Vincent Doyle:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Not applicable

References

1. Eccles MP, Mittman BS. Welcome to implementation science. *Implement Sci*. 2006;1(1):1. https://doi.org/10.1186/1748-5908-1-1.
2. Adapting NICE guidelines | NICE International | What we do | About. NICE. Accessed January 3, 2024. https://www.nice.org.uk/about/what-we-do/nice-international/adapting-nice-guidelines.
3. Wang C, Ong J, Wang C, Ong H, Cheng R, Ong D. Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. *Ann Biomed Eng*. 2024;52(5):1115–1118. https://doi.org/10.1007/s10439-023-03327-6.

4. Nadarzynski T, Knights N, Husbands D, et al. Achieving health equity through conversational AI: a roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digit Health*. 2024;3(5). https://doi.org/10.1371/journal.pdig.0000492 e0000492.
5. NICE. *NICE Strategy 2021 to 2026*. NICE; 2021. https://www.nice.org.uk/Media/Default/Get-involved/Meetings-In-Public/Public-board-meetings/Mar-24-pbm-NICE-strategy-2021-2026.pdf.
6. Activity In The NHS. The King's Fund. Accessed August 7, 2024. https://www.kingsfund.org.uk/insight-and-analysis/data-and-charts/NHS-activity-nutshell.
7. NHS workforce: Record numbers of doctors and nurses in NHS – Department of Health and Social Care Media Centre. Accessed August 7, 2024. https://healthmedia.blog.gov.uk/2023/04/27/nhs-workforce-record-numbers-of-doctors-and-nurses-in-nhs/.
8. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. *BMJ Lead. Published Online*. 2023. https://doi.org/10.1136/leader-2023-000797. (leader).
9. Morley J, DeVito NJ, Zhang J. Generative AI for medical research. *BMJ*. 2023;382. https://doi.org/10.1136/bmj.p1551 p1551.
10. What Is ChatGPT Doing . . . and Why Does It Work? February 14, 2023. Accessed January 3, 2024. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.
11. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Published online. 2021. doi:10.48550/arXiv.2005.11401.
12. ChatNice - UK's medical AI chatbot. Accessed November 28, 2023. https://gloriamacia.wixsite.com/chatnice.
13. Build a multi-tenant chatbot with RAG using Amazon Bedrock and Amazon EKS | Containers. 2023. Accessed August 8, 2024. https://aws.amazon.com/blogs/containers/build-a-multi-tenant-chatbot-with-rag-using-amazon-bedrock-and-amazon-eks/.
14. LangChain. Accessed. 2024. https://www.langchain.com/.
15. Liu J. LlamaIndex. Published online 2022. doi:10.5281/zenodo.1234.
16. Introducing GPT-4o and more tools to ChatGPT free users. Accessed August 8, 2024. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.
17. New embedding models and API updates. Accessed August 8, 2024. https://openai.com/index/new-embedding-models-and-api-updates/.
18. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. Published online February 28; 2017. doi:10.48550/arXiv.1702.08734.
19. Sharma V, Raman V. A reliable knowledge processing framework for combustion science using foundation models. *Energy AI.*. 2024;16. https://doi.org/10.1016/j.egyai.2024.100365 100365.
20. Using LLM-as-a-judge for an automated and versatile evaluation - Hugging Face Open-Source AI Cookbook. Accessed August 7, 2024. https://huggingface.co/learn/cookbook/en/llm_judge.
21. Thakur AS, Choudhary K, Ramayapally VS, Vaidyanathan S, Hupkes D. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. Published online July 1, 2024. doi:10.48550/arXiv.2406.12624.
22. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. Curran Associates Inc.; 2024:46595-46623.
23. Overview | Caesarean birth | Guidance | NICE. March 31, 2021. Accessed November 28, 20https://www.nice.org.uk/guidance/ng192.
24. The Importance of Diversity and Inclusion in the Healthcare Workforce - PMC. Accessed January 3, 20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7387183/.
25. Privacy in the Time of Language Models | Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. Accessed January 3, 2024. https://dl.acm.org/doi/abs/10.1145/3539597.3575792.
26. ChatGPT Not Ready for Prime Time for Medication Queries. Accessed January 3, 2024. https://www.pharmacypracticenews.com/Pharmacy-Technology-Report/Article/12-23/ChatGPT-Not-Ready-for-Prime-Time-for-Medication-Queries/72307.