ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2026.2333

Comparing five generative AI chatbots' answers to LLM-generated clinical questions with medical information scientists' evidence summaries

Mallory N. Blasingame; Taneya Y. Koonce; Annette M. Williams; Jing Su; Dario A. Giuse; Poppy A. Krump; Nunzia B. Giuse

See end of article for authors' affiliations.

Objective: To compare answers to clinical questions between five publicly available large language model (LLM) chatbots and information scientists.

Methods: LLMs were prompted to provide 45 PICO (patient, intervention, comparison, outcome) questions addressing treatment, prognosis, and etiology. Each question was answered by a medical information scientist and submitted to five LLM tools: ChatGPT, Gemini, Copilot, DeepSeek, and Grok-3. Key elements from the answers provided were used by pairs of information scientists to label each LLM answer as in Total Alignment, Partial Alignment, or No Alignment with the information scientist. The Partial Alignment answers were also analyzed for the inclusion of additional information.

Results: The entire LLM set of answers, 225 in total, were assessed as being in Total Alignment 20.9% of the time (n=47), in Partial Alignment 78.7% of the time (n=177), and in No Alignment 0.4% of the time (n=1). Kruskal-Wallis testing found no significant performance difference in alignment ratings between the five chatbots (p=0.46). An analysis of the partially aligned answers found a significant difference in the number of additional elements provided by the information scientists versus the chatbots per Wilcoxon-Rank Sum testing (p=0.02).

Discussion: Five chatbots did not differ significantly in their alignment with information scientists' evidence summaries. The analysis of partially aligned answers found both chatbots and information scientists included additional information, with information scientists doing so significantly more often. An important next step will be to assess the additional information, both from the chatbots and the information scientists for validity and relevance.

Keywords: Large Language Models; LLMs; generative AI; chatbots; artificial intelligence; evidence synthesis; library science; information science; biomedical informatics

See end of article for supplemental content.

INTRODUCTION

Generative artificial intelligence (AI) tools are increasingly embedded into the systems and workflows used by experts and the general public to search for health information. In 2024, a Kaiser Family Foundation poll of 2,428 U.S. adults found that roughly 1 in 6 (17%) respondents used AI chatbots at least monthly to seek out health information [1]. Even when searchers do not directly query a generative AI tool, they may increasingly encounter large language model (LLM)-generated answers as they search for health information on the web [2] or in proprietary literature databases [3,4]. Google and Microsoft Bing are now providing generative AI answers in search results, including responses to medical queries [5]. National Center for Health Statistics data from 2022 revealed that 58.5% of adults surveyed had looked for health information on the Internet in the past year [9], and Google reported in 2025 receiving "100s of millions" of health-related searches per day [10]. Thus, it is likely that many people are encountering generative AI answers to their everyday health inquiries. These answers can provide users with quick, easy access to synthesized information, which may be useful for guiding conversations with clinical providers [6] but may also pose unforeseen risks [7,8].

Furthermore, with the increased integration of LLMs within the realm of searching and summarizing clinical information, evaluating how highly-used models' answers compare to those of the highly trained, trusted medical librarians/information scientists who commonly perform

DOI: dx.doi.org/10.5195/jmla.2026.2333

these tasks may greatly help our understanding of their utility, limitations, and risks. Previous studies have investigated LLMs' performance for answering medical questions using a variety of study designs and evaluation dimensions [11] and assessed their ability to aid with steps of the evidence synthesis process including search strategy development [12–14] and systematic review tasks such as citation screening and data extraction [15,16]. However, to fully understand the implications of LLMs to the medical library profession, additional investigation is needed into how AI chatbots' answers to medical questions compare to librarian-generated evidence summaries.

To address this knowledge gap, our team at the Vanderbilt University Medical Center (VUMC) Center for Knowledge Management (CKM) has embarked on a series of studies to investigate generative AI in the context of medical information sciences. Our team is composed of professional information scientists credentialed in medical librarianship and, in most instances, in one or more additional health sciences disciplines, e.g., medicine, public health, bioinformatics. In a previous study, we assessed the performance of VUMC's internally managed version of GPT-4 (aiChat) [17], using medical information scientists' evidence summaries as the gold standard for comparison [18]. In that initial study, we compared aiChat answers with summaries our team previously developed in response to a VUMC-proprietary set of questions received during rounds or from an electronic health record-linked message basket. The study revealed that 83.3% of aiChat responses included all the elements from the information scientist summaries that were identified as being most critical for answering the questions, while also reflecting known limitations of generative AI tools, including fabrication of references. We additionally observed that the well-organized, consistent formatting of the summaries was a strength of the LLM-generated answers. In a second study comparing search strategies generated by three publicly available LLMs with information scientists' expert searches, our team found the AI chatbots were able to generate Boolean search queries but missed many relevant keywords and often included inaccurate controlled vocabulary terms [19].

As part of our contribution and ongoing commitment to increase librarians' understanding on how to best integrate AI in our profession, our team has designed a series of evaluations that will, step by step, explore a variety of LLM-related research questions, thus increasing our overall knowledge on generative AI. The use in our first study of institutional proprietary data limited the study to a single large language model, aiChat. The current study extends the investigation of LLMs' performance in clinical question answering to additional tools. We aimed to build on the findings of the previous two studies by conducting a prospective, in-depth, detailed comparison between five generative AI chatbots' answers to clinical questions and medical information scientists' synthesized evidence summaries, focusing on tools that are widely available and likely to be used by the public. To avoid using proprietary questions originating from our medical center's clinical work environment in these publicly available systems, we prompted the chatbots to create the questions for the study, and information scientists developed new synthesized evidence summaries for each question to compare with the LLMs' answers. As the summaries created by the information scientists were not reviewed and validated by clinical experts, we did not consider them to be a "reference standard" for assessing accuracy but rather compared them one-to-one with the chatbots' answers. The study specifically investigated the following questions:

1) How do the answers provided by five publicly available LLMs compare to those of medical information scientists?
2) Are there differences in how the LLM answers compare with information scientist answers when the AI tools answer their own generated questions?

METHODS

The study received a non-human subjects research determination from the VUMC Institutional Review Board (IRB #241743). The reporting of this study follows the Chatbot Assessment Reporting Tool (CHART) guidelines [20,21]; the CHART Methodological Diagram and Checklist can be viewed in Appendix A.

Generative AI Chatbots

When the study was initially conceived in October 2024, ChatGPT (https://chatgpt.com/), Google Gemini (https://gemini.google.com/), and Microsoft Copilot (https://copilot.microsoft.com/) were selected for investigation based on their reported frequency of use in medical literature to date [22], and, in the cases of Gemini and Copilot, their increasing integration into highly used public search engines. In early 2025, as study activities were ongoing, the release of DeepSeek (https://chat.deepseek.com) and Grok-3 (https://grok.com/) for public use sparked a great deal of interest; thus, we decided to add them to the analysis to explore any differences in performance with these newer models. The free, public, web-based versions were used; in Google Gemini, DeepSeek, and Grok-3, logging into a free personal account was required. The versions used were the most current freely available, base models for each tool at the time: ChatGPT-4o [23], Gemini 2.0 Flash [24], DeepSeek R1 [25], and Grok-3 [26] with web search enabled; the particular model of Copilot was not specified in the system. DeepSeek defines itself as an open-source LLM; all other models used in this study are closed-source. Though Microsoft Copilot uses OpenAI's GPT as one of the base model options, given the differences in

DOI: dx.doi.org/10.5195/jmla.2026.2333

Microsoft's integration of the GPT model, combined with Microsoft Copilot's auto-routing to select a model based on the context of the user's input, the two LLMs were considered as independent for this investigation [27].

Prompt Engineering

Three prompts were used in the study: 1) a prompt to obtain the questions from the generative AI chatbots; 2) a prompt to obtain the *answers* from the LLMs; and 3) a prompt to submit each information scientist and chatbot answer to the generative AI tools to obtain lists of key elements for comparison. The prompt for obtaining the answers was reused from our previous study [18]. The other two prompts were newly created for this study by the co-authors, who include individuals with formal training and expertise in information sciences, medicine, public health, and biomedical informatics. The COSTAR framework [28] was followed, with each prompt including a section on Context, Objective, Style, Tone, Audience, and Response. When possible, elements of the original prompt were reused, with adjustments made to tailor the prompt to the specific task. The prompts were submitted to the tools for testing and revised as needed. In all cases, a new session was started with the chatbot for each individual prompt submitted.

Obtaining the Questions

In our previous study [18], we used questions received from clinicians based on actual clinical encounters; these questions are proprietary to our institution and thus could only be used with our organization's internally managed AI tool. For this study that uses publicly available LLMs, we intentionally used the chatbots to create non-proprietary clinical questions that could be input into the tools and publicly shared (see Appendix B). In evidence-based medicine, the use of well-structured questions can enhance precision and aid in establishing discrete concepts for search strategy formulation and information retrieval [29]. Since the 1990s, the PICO (patient, intervention, comparison, outcome) framework has been used by clinicians and information scientists to guide evidence searching, filtering, and selection in response to both clinical questions and inquiries from members of the lay public [30]. Thus, we leveraged the chatbots' ability to quickly and efficiently generate PICO-structured questions for use in this study. Each question was subsequently reviewed by an information scientist with formal education and training in medicine to ensure all questions generated were medically plausible, i.e., that the questions made medical and logical sense.

Forty-five questions were obtained in November 2024 by prompting ChatGPT, Google Gemini, and Microsoft Copilot to each provide five treatment questions, five prognosis questions, and five etiology questions in PICO format (Appendix B). DeepSeek and Grok-3 were not used to generate questions, as the tools were not yet widely available for public use. The treatment, etiology, and prognosis question categories were selected based on the most common types of questions our team has received in our history of providing clinical evidence services [18]. After initial prompt testing revealed that the tools tended to provide questions focused on treatment even when asked for prognosis or etiology questions, definitions of prognosis and etiology were added to the prompts for these categories (Appendix C). A definition of treatment was not added to the prompt, as the tools were able to provide questions in this category without the need for one.

The questions were obtained by two information scientists. When duplicate or non-medically plausible questions were generated, they were removed from the question pool and the tool that provided the question was prompted for an additional question as part of the same session. For example, one generated question asked about a patient population with Type 2 diabetes whose HbA1c level was within a "normal range at diagnosis"; given elevated HbA1c is a standard diagnostic criterion for Type 2 diabetes [31], we excluded this question and prompted the chatbot to provide another.

Question Assignment

Questions were assigned to four information scientists based on their effort assigned to the project. The questions were initially randomly assigned with stratification by question source and category. As efforts shifted throughout the project period, five questions were reassigned.

Developing the Information Scientist Evidence Summaries

Our team's standard practices for developing synthesized evidence summaries were followed. As often done at our center, information scientists individually developed comprehensive PubMed search strategies for each question and then met as a group to review the searches and provide feedback on areas for refinement. Once the searches were finalized, the information scientists completed evidence summaries for their questions, following the template used by our team for answering real-world evidence queries. The template reflected changes adopted by the team after our previous AI study [18] and includes the following sections: 1) an introduction, including a brief summation of findings, characterization of the state of the literature, and definition of key topics; 2) a summary of selected literature, including the design, publication year, aims, and results for each selected study; and 3) conclusion, with the "bottom line" of findings from the literature and a brief summary of strengths and limitations. After each summary was completed, it was stored in REDCap [32,33].

DOI: dx.doi.org/10.5195/jmla.2026.2333

Obtaining Answers from the Generative AI Tools

As the information scientists finalized each summary, a senior information scientist not involved in evidence synthesis submitted the corresponding question to each chatbot using a standardized prompt (Appendix D). In brief, the prompt asked the tools to provide an evidence summary in response to the provided clinical question in "the role of a medical librarian," with the answer limited to the information available up to the date the information scientist completed the summary. All information scientist summaries were completed and answers captured from the generative AI tools between January and April 2025. Each response was stored in full in the study database, along with references and any hyperlinks included with the response. Both the information scientist and generative AI answers are available in the complete study dataset (see Data Availability Statement).

Comparing the Information Scientist Evidence Summaries with Generative AI Answers

In our research assessments of generative AI tools, CKM recognized, although limitations were observed in the ability to verify LLM-provided references, there was value in the tool's capability to clearly and effectively summarize evidence in a well-organized written format and, when specified, organize the information into easy-to-understand key elements. Leveraging on this understanding and with the intent to remove human subjectivity, for the current study, the team decided to have each of the five tools automatically generate (Appendix E) the key elements for the 45 LLM-generated summaries and the 45 summaries generated by the information scientists. Each set of key elements from all the summaries, whether generated by the tools or written by the information scientists, was subsequently reviewed by information scientists to ensure the key elements accurately reflected the summary.

Because of the lack of direct involvement and interaction with clinicians, a decision was also made not to label the information scientists' summaries as "reference standards." Without the opportunity for consultation, we cannot assume that information scientists and clinical experts will agree on the answers to clinical questions [34]. Given the above consideration, agreement and disagreement among the key elements was used to determine the level of alignment between the answers of the LLMs and information scientists; the analysis was conducted by four unblinded information scientists. All key elements are available in the study data deposited to the Open Science Framework. A pair of information scientists reviewed and gathered consensus on whether the information included in the key elements from the answers of each of the five models – ChatGPT, Gemini, Copilot, DeepSeek, and Grok-3 – was a) totally aligned, b) partially aligned, or c) not aligned with the information expressed by the key elements from the information scientists' answers. In each of the above instances, one of the information scientists in the pair was the author of the summary compared.

We used the concept of *Total Alignment* when the information expressed by the key elements in each of the answers being compared was judged as being the same. Although the lists of key elements were numbered, the summaries did not need to have the same number of elements to be considered in Total Alignment. For example, a single key element from the information scientist's response could be determined to align with two or more elements from the chatbot's response, and vice versa. Additionally, differences in wording or data cited that represented the same concept or viewpoint were counted as in alignment. *Partial Alignment* included four sub-categories: 1) All key elements expressing the same concepts were in agreement, but the *tool's* answer included additional key elements; 2) All key elements expressing the same concepts were in agreement, but the *information scientist's* answer included additional key elements; 3) All key elements expressing the same concepts were in agreement, but *both* answers included additional key elements; or 4) The information scientist and LLM agreed on some but not all key elements expressing the same concepts, and one or both answers may have included additional key elements. For categories 1-3, we tabulated how many additional key elements were provided, meaning concepts not found in the answers used for comparison. *No Alignment* indicated that none of the key elements agreed.

Sample Size Determination

The sample size was informed by the need to obtain multiple questions from each of the three tools we originally intended to study (ChatGPT, Gemini, and Copilot) and the three categories selected for the questions (treatment, prognosis, and etiology). Consideration was also given to the estimated time (historically averaging about 8 hours per summary [35]) needed by the information scientists to generate summaries, a consideration largely dictated by their availability. All of this brought us to the final determination that using a minimum of five questions for each of the question categories (total of 15 questions) for each of the three tools could give us adequate data for our study, resulting in a final sample of 45 questions. A formal sample size calculation was not conducted.

Data Analysis

Descriptive statistics were used to report the frequency and proportion of responses assessed as in Total Alignment, Partial Alignment, or No Alignment for each of the tools in comparison with the information scientist responses, and to characterize the number and percentage of each type of Partial Alignment by tool. The average word counts for the narrative text, excluding the list of

DOI: dx.doi.org/10.5195/jmla.2026.2333

cited references, from the information scientist and chatbot answers were also calculated. A Kruskal-Wallis test was used to assess whether there was a significant difference in the five generative AI tools' alignment with the information scientists' responses, and in the types of partial alignment across the five tools. This test was selected as the comparison focused on multiple independent groups and the data were nonparametric.

For the partially aligned answers in categories 1-3 above, the median number and mean proportion of key elements identified as "additional" from the information scientist and LLM were calculated. A Wilcoxon Rank-Sum test was used to compare the two groups of data on the number of additional key elements from the information scientist summaries versus the chatbot summaries as another measure of content differences between the information scientist and LLM summaries. In a sub-analysis focused on the three generative AI tools that provided the PICO questions (ChatGPT, Gemini, and Copilot), a Friedman statistic test was used to assess whether there were any significant differences in each of the three tools' alignment based on whether they were answering their own questions or those provided by the other two chatbots. The Friedman statistic test was chosen due to the use of dependent groups and rank-based data. All statistical analyses were conducted in GraphPad Prism, and visualizations were created in Flourish and Microsoft Excel.

RESULTS

In total, 53 PICO questions were generated by ChatGPT, Gemini, and Copilot. During the process of generating each initial set of 15 questions from the three tools, eight questions were removed from the pool after review and eight new questions generated, resulting in the final set of 45 questions. Reasons for question exclusion included determination that the question was medically implausible (n=3; 2 from ChatGPT, 1 from Gemini), that the question was a duplicate of another question already in the pool (n=4; 2 from Copilot, 2 from ChatGPT), or that the question did not align with the requested category (n=1 from ChatGPT).

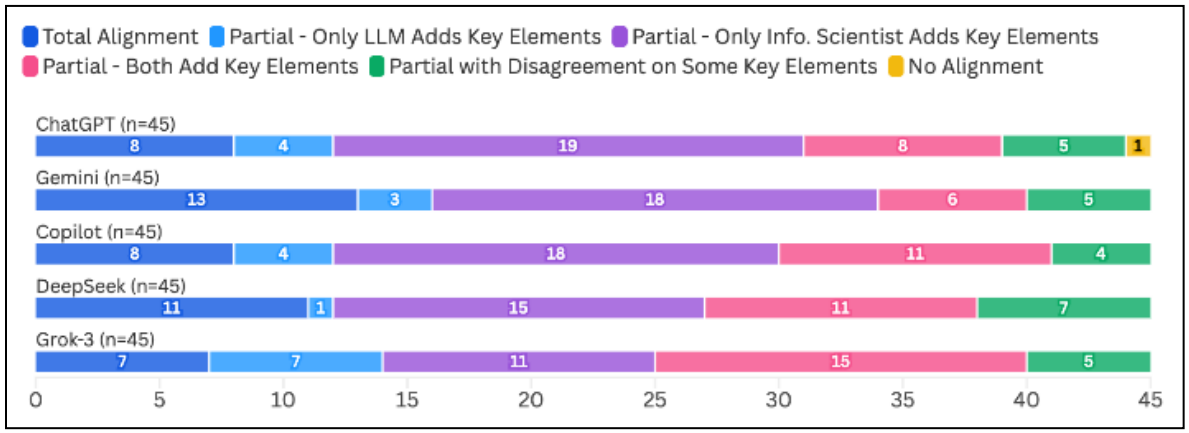
Alignment between the Information Scientist and Generative AI Responses

Across the 225 answers generated to the 45 questions by the five generative AI tools, 47 (20.9%) were in Total Alignment with the information scientist's response, 177 (78.7%) were in Partial Alignment, and one (0.4%) was assessed as having No Alignment. The response with No Alignment was from ChatGPT. Gemini had the highest frequency of Total Alignment ratings (13/45; 28.9%), while Grok-3 had the lowest frequency of answers in Total Alignment (7/45; 15.6%). A Kruskal-Wallis test revealed no significant differences in the alignment ratings between the five tools (p=0.46). The full alignment ratings for each chatbot can be viewed in Figure 1.

Analysis of Partially Aligned Responses

In the subset of 177 responses that we labeled as in Partial Alignment, 151 (85.3%) had all key elements expressing the same concepts in agreement, but the answers from the generative AI tool (n=19), the information scientist (n=81), or both (n=51) contained additional unique elements. The remaining 26 (14.7%) answers labeled as Partial Alignment agreed only on some of the key elements expressing the same concepts; in all cases in this category, there was only one concept identified as being in disagreement. Grok-3 had the highest frequency of responses in which only the LLM included additional key elements (7/45; 15.6%) or both the tool and the information scientist provided additional key elements (15/45; 33.3%). ChatGPT had the most answers for which only the information scientist included additional key

Figure 1 Alignment of Five LLMs' Answers to 45 PICO Questions Compared to Information Scientists' Responses (N=225 LLM answers)



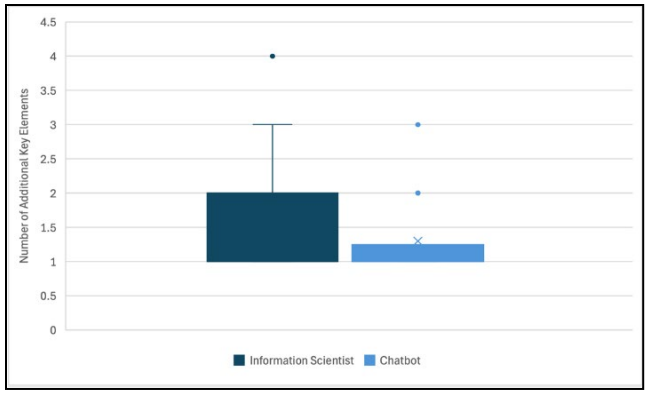
DOI: dx.doi.org/10.5195/jmla.2026.2333

elements (19/45; 42.2%), and DeepSeek had the most answers that did not agree on all key elements in common with the information scientist's answer (7/45; 15.6%). However, a Kruskal-Wallis analysis found no significant differences in type of Partial Alignment among the five tools (p=0.78). Full Partial Alignment results by tool can be viewed in Figure 1.

Analysis of Additional Key Elements

The median number of total elements from the *chatbot* answers providing additional information (total of 70 answers) was 5.5 (range: 3-8). There was a median of 1 (range: 1-3) additional key element per answer. The median number of total elements from the *information scientists'* answers with additional key elements (total of 132) was 5 (range: 3-10). There was a median of 1 (range: 1-4) additional key element per answer. The distribution of the number of additional key elements identified in each information scientist and chatbot answer in this category is shown in Figure 2. A Wilcoxon Rank-Sum test found a significant difference when comparing the groups of additional key elements generated for the chatbot summaries versus the information scientist summaries (p=0.02).

Figure 2 Distribution of Additional Key Elements from the Information Scientist (n=132) and Chatbot Answers (n=70)



Alignment by Question Source

In the sub-analysis limited to the three LLMs that provided the questions (ChatGPT, Gemini, and Copilot), all three of the chatbots had the highest frequency of Total Alignment ratings when answering Copilot questions (Figure 3). However, per Friedman statistic testing, no significant differences in alignment ratings were found for any of the three tools when comparing the alignment ratings of their answers to their own questions against the alignment ratings of their answers to the other two tools' questions (p=0.24 for ChatGPT, 0.88 for Gemini, and 0.56 for Copilot). Full alignment ratings by question source for these three tools can be viewed in Figure 3.

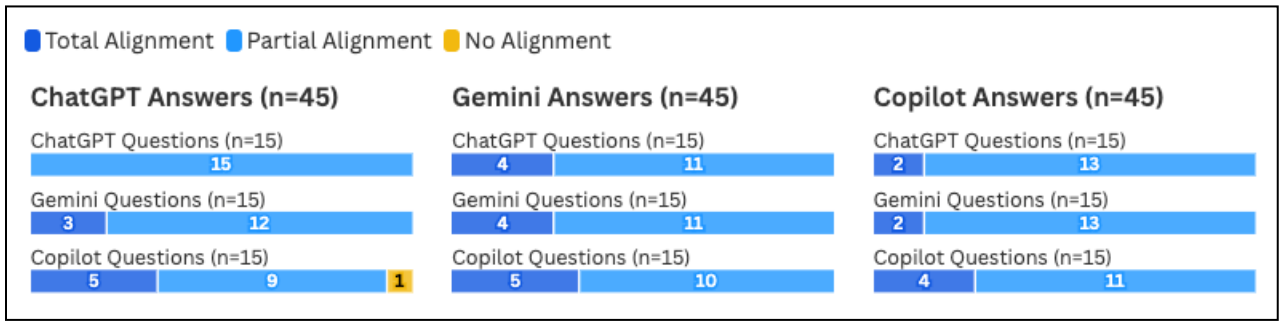
Descriptive Characteristics of the Information Scientist and Generative AI Summaries

On average, the information scientists' summaries were longer than the generative AI answers, with a mean of 1,668 words observed for the information scientists' answers compared with average word counts ranging from 286 (Copilot) to 497 words (Grok-3) for the generative AI tools' answers. The generative AI tools commonly included direct links to references cited in the summaries (100% of ChatGPT answers had linked references, 91% of Gemini answers, 84% of Copilot answers, and 60% of Grok-3 answers), except for DeepSeek, which did not provide direct links to cited references. Consistent with our team's standard practice, all 45 information scientist summaries included full references with hyperlinks.

DISCUSSION

This study is the second in what we envision being a step-by-step series to clearly investigate all elements of AI performance as applicable to the functional work of information scientists. Our comparison between information scientists' synthesized evidence summaries and five generative AI chatbots revealed that in most instances (20.9% total and 78.7% partial), there is a high level of alignment among the summaries evaluated, even

Figure 3 Alignment between Information Scientist and LLM by Individual Chatbot (ChatGPT, Gemini, and Copilot)



DOI: dx.doi.org/10.5195/jmla.2026.2333

though an in-depth analysis of the partial answers shows that information scientists are more likely to include additional information (key elements) in their answers. Central to this study is the reporting of additional key elements observed in either the large language models' or the information scientists' summary responses and the analysis comparing information scientists' answers to five distinct models. No significant differences were found in the degree to which the five generative AI tools' responses aligned with the information scientists' answers, and no impact on performance was observed when the tools answered their self-generated questions, suggesting that the study results were not influenced by whether the questions were based on the LLM's own training data.

It is worth noting that other studies have found significant differences in LLMs' answers to clinical questions, although variation in models, question types, and overall methodology makes it difficult to directly compare the existing literature [11]. For example, Lin and colleagues (2025) compared several ChatGPT, Gemini, and Copilot models' answers to questions on postmenopausal osteoporosis with guideline recommendations and found that ChatGPT-4o's answers were significantly more accurate than other models [36]. Flaharty et al. (2024) investigated several open- and closed-source models' performance (including Gemini and ChatGPT-3.5 and 4) in answering genetic questions and found that ChatGPT-4 had the highest performance in terms of correctly identifying a genetic condition, with significant differences observed in several models' performance depending on whether the question was asked in medical or lay language [37].

We conducted an in-depth analysis of the only "non-alignment" answer in the study (generated by ChatGPT). The answer addressed a question about the use of early ischemic changes for prediction of "functional outcomes" in acute ischemic stroke patients. In comparing the two responses, we noted that the ChatGPT answer included a reference supporting the opposite viewpoint from the information scientist's answer. The answer, although not aligned with the information scientist, was reported as one of the multiple viewpoints provided by the other tools, giving us a moment of pause and reinforcing our belief that all additional elements provided by LLMs deserve further investigation. Our study also found the information scientists' answers more commonly provided additional information than the generative AI chatbots' responses. This difference may be explained in part by the fact that information scientists' evidence summaries were longer on average than the LLMs'. It is however notable that, despite their shorter length, 70/225 (31.1%) generative AI summaries were found to provide key elements not present in the information scientists' summaries. This finding presents a potential for chatbots to serve a complementary role in evidence synthesis by surfacing additional supporting information and/or alternate viewpoints for the information scientist to investigate, verify, and consider for inclusion.

As noted previously, we intentionally did not make a judgment on which answer was correct but rather characterized the answers in terms of alignment between the chatbot and human, with no true "reference standard." Evidence synthesis is a complex task often involving interaction through the reference interview and ongoing consultation between the subject matter expert and information professional. Hripcsak and Wilcox (2002), in their discussion of reference standards for evaluations of informatics systems, state, "For more complex reasoning tasks, experts are needed to judge the appropriateness of system responses," and an answer may be assessed by a clinical expert as "appropriate even if it matches none of the comparison responses (e.g., a reasonable medication alternative)" [38]. When a true reference standard is lacking, answers prepared by a human can be compared with answers from the system for "similarity" instead of "performance" [38], consistent with the approach used in our study.

The need to carefully review LLM responses for accuracy is well-established and critical to their use [39]. As evidenced by the non-aligned example, chatbots may also introduce additional information from legitimate sources that cannot be assumed to be incorrect without further examination. Humans may also miss relevant information or have differing assessments of relevance depending on their role or area of specialization [37,40]. The focus on alignment and additional information provided by both human and chatbot in this study highlights the potential for LLMs to serve as an aid in a framework where the information scientist prepares an answer to an evidence inquiry but uses the chatbot to identify additional information for further investigation and inclusion in her citations and summaries.

These findings build on our approach, throughout our series of studies, of applying a "growth mindset" to the investigation of how AI can be applied to our workflows as information scientists [40]. Similar to the process of consulting a colleague and deciding if and how to incorporate their ideas and feedback, consulting a chatbot can help stimulate our thinking and uncover previous knowledge gaps. By identifying LLMs' areas of strength and applying them to our work as information scientists, we can begin to truly partner with AI to enhance our performance beyond only productivity gains [40].

Limitations

Limitations of the study include not assessing the chatbot answers for potential harm or misleading statements. The next planned phase – examining the additional information from the chatbots – will be an important step for further elucidating any potential unfounded

DOI: dx.doi.org/10.5195/jmla.2026.2333

information that may have been included in their responses.

Additionally, the information scientist who answered each question was included in the pair of reviewers who assessed the chatbots' answers. While this was done intentionally to leverage the information scientist's knowledge of the topic, it is possible this approach introduced cognitive bias.

As we wanted to create an optimal scenario for question answering, we intentionally had the chatbots generate PICO-formatted questions. However, the use of LLM-generated PICO questions may limit the study's generalizability to actual clinical scenarios. While the PICO format provides a useful and well-established framework to prompt LLMs to generate questions in alignment with prompt engineering techniques [41], it is not always characteristic of questions that arise in fast-track clinical settings. As illustrated by a study from Huang and colleagues (2006), clinicians' questions are often missing one or more of the PICO elements [42]. Nevertheless, the PICO format is one of the most common mechanisms librarians use to fully research and investigate questions they receive from their users, making this type of exploration highly relevant for understanding the potential role of AI usage in medical librarianship.

Four of the chatbots included in the study are closed-source models; thus, reproducibility is limited by the lack of transparency into the models' training sets and design. We did not conduct a formal investigation into reproducibility of chatbot answers. We also acknowledge that as time passes and models are updated, reproducibility may be limited as the chatbots have access to more data, due to both the emergence of new knowledge and additions to the LLMs' training sets. Stanford's Holistic Evaluation of Language Models for Medical Tasks (medHELM) Leaderboard demonstrates that performance on medical domain scenarios varies between model versions, with Gemini, for example, showing improved accuracy with 2.0 Flash versus 1.5 Pro [43]. We also acknowledge that the study, by design, was limited to five large language models and did not assess the performance of other models such as Claude or Perplexity. It is possible different results would have been observed for other models due to variations in training data or model design.

Finally, the comparison of the LLM and information scientist answers relied on the key elements selected by each tool. The information scientists agreed that the elements selected were appropriate, but we did not formally assess whether the lists of key elements for their summaries were consistent across all five tools.

Future Directions and Conclusions

In this study, submitting PICO questions to publicly available generative AI chatbots using a standardized prompt revealed that chatbots both added and missed key information relative to information scientists' answers. These findings provide insight into LLMs' capabilities and their potential utility to complement information scientists' evidence synthesis processes. On a practical level, chatbots have the potential to supplement information scientists' expertise in searching and filtering the literature by serving as an additional information source for consultation, as well as aiding with quick summarization and extraction of key elements of the narrative response for sharing with end users such as clinicians. We recognize that, given the rapid evolution of the field and emergence of new models, ongoing validation of LLM tools will be paramount. Full adoption in healthcare will additionally require a complete transparency of the data sources for evidence verification; issues of privacy and security will need to be fully addressed as well.

A critical next phase of this investigation will be to examine all the additional information provided by the chatbots to determine whether it can be validated and supported by the literature, as this will be especially essential for fully understanding the implications for their use by the public. When clinicians review an answer with unfamiliar or unexpected information, they can ask the information scientist and now chatbots for additional clarifying data and, in most instances, they are able to discern the added value of the information they receive. Still, it is important to note that chatbot users may be subject to what is commonly known as automation bias, the propensity to trust information from automated systems without further investigation [40]. Thus, the need to assess the validity of the entire content provided by the answers of the tools as well as the information scientists. Additionally, further investigation into how the public understands and interacts with LLMs to find health information may be needed, as studies have found that their performance in answering health-related queries was rated higher when prompted directly by researchers than when prompted by members of the public [44,45].

Although not formally analyzed in this study, we observed that four of the LLMs (all except DeepSeek) commonly included direct links to supporting articles and websites in their answers. In contrast, the answers we previously analyzed from VUMC's aiChat tool only included in-text references with no direct links, many of which could not be verified to exist [18]. Being able, with external and newer versions of the models that provide web search features, to directly access and consult cited references is an improvement in facilitating users' ability to verify chatbots' answers. A future planned analysis will conduct a more in-depth examination of the references cited by all five LLMs. As the adoption of AI becomes

DOI: dx.doi.org/10.5195/jmla.2026.2333

more widely accepted in healthcare, a blinded study assessing clinicians' preferences between information scientist and chatbot answers and a study comparing LLMs' answers to a gold standard vetted by expert clinicians could offer additional insight to clinical librarians engaged in evidence provision.

This study constitutes a key step in our series of investigations to understand large language models' utility in medical information sciences by revealing a lack of significant differences in five chatbots' alignment with medical information scientists' answers and demonstrating common overlap in key elements of the responses. While information scientists were significantly more likely to contribute additional information not included in the chatbots' answers, the fact that the LLMs also provided additional information not included by the information scientist suggests a potential framework in which these tools could play a consultative role in medical evidence synthesis.

FUNDING STATEMENT

The REDCap database, used in this study for data collection and storage, is supported by CTSA award UL1TR000445 from the National Center for Advancing Translational Sciences.

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS STATEMENT

Mallory N. Blasingame: Methodology; investigation; data curation; formal analysis; visualization; writing–original draft; writing–review and editing. Taneya Y. Koonce: Methodology; investigation; data curation; formal analysis; writing–original draft; writing–review and editing. Annette M. Williams: Methodology; investigation; data curation; writing–review and editing. Jing Su: Methodology; investigation; visualization; writing–review and editing. Dario A. Giuse: Methodology; investigation; writing original draft; writing–review and editing. Poppy A. Krump: Methodology; investigation; writing–review and editing. Nunzia B. Giuse: Conceptualization; methodology; investigation; formal analysis; visualization; writing–original draft; writing–review and editing; supervision.

ACKNOWLEDGEMENTS

This research was developed through the training and support provided by the Medical Library Association's Research Training Institute (RTI). The authors would like to acknowledge Spencer DesAutels for his assistance with data visualization.

DATA AVAILABILITY STATEMENT

Data associated with this study are available from the Open Science Framework at https://doi.org/10.17605/OSF.IO/BJ9TH.

REFERENCES

1. Presiado M, Montero A, Lopes L, Hamel L. KFF health misinformation tracking poll: artificial intelligence and health information [Internet]. Kaiser Family Foundation; 15 Aug 2024 [cited 23 Sept 2025]. https://www.kff.org/health-misinformation-and-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/.
2. Chapekis A, Lieb A, Shah S, Smith A. What web browsing data tells us about how AI appears online [Internet]. Pew Research Center; 23 May 2025 [cited 23 Sept 2025]. https://www.pewresearch.org/data-labs/2025/05/23/what-web-browsing-data-tells-us-about-how-ai-appears-online/.
3. Taylor J, Dagan K, Youngberg M, Kaufman T, Radding J. A survey of AI tools in library tech: accelerating into and unlocking streamlined enhanced convenient empowering game-changers. J Electron Resour Librariansh.2025 May;1–14. DOI: https://doi.org/10.1080/1941126X.2025.2497738.
4. Livingston L, Featherstone-Uwague A, Barry A, Barretto K, Morey T, Herrmannova D, Avula V. Reproducible generative artificial intelligence evaluation for health care: a clinician-in-the-loop approach. JAMIA Open. 2025 June;8(3):ooaf054. DOI: https://doi.org/10.1093/jamiaopen/ooaf054.
5. Yau JY, Saadat S, Hsu E, Murphy LS, Roh JS, Suchard J, Tapia A, Wiechmann W, Langdorf MI. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res. 2024 Nov 4;26:e60291. DOI: https://doi.org/10.2196/60291.
6. Sundar KR. When patients arrive with answers. JAMA. 2025 Aug;334(8):672-3. DOI: https://doi.org/10.1001/jama.2025.10678.
7. Ashraf AR, Mackey TK, Fittler A. Search engines and generative artificial intelligence integration: public health risks and recommendations to safeguard consumers online. JMIR Public Health Surveill. 2024 Mar;10:e53086. DOI: https://doi.org/10.2196/53086.
8. Eichenberger A, Thielke S, Van Buskirk A. A case of bromism influenced by use of artificial intelligence. Ann Intern Med Clin Cases. 2025 Aug;4(8):e241260. DOI: https://doi.org/10.7326/aimcc.2024.1260.
9. Wang X, Cohen RA. Health information technology use among adults: United States, July-December 2022 [Internet]. NCHS Data Brief No. 482. Hyattsville, MD: National Center for Health Statistics; 2023 [cited 23 Sept 2025]. https://www.cdc.gov/nchs/products/databriefs/db482.htm.
10. DeSalvo K. Google's impact on health [Internet]. Mountain View, CA: Google; Feb 2025 [cited 17 Oct 2025].

DOI: dx.doi.org/10.5195/jmla.2026.2333

https://services.google.com/fh/files/misc/googles_health_impact.pdf.

11. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS, Hong HJ, Kashyap M, Chaurasia AR, Shah NR, Singh K, Tazbaz T, Milstein A, Pfeffer MA, Shah NH. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025 Jan 28;333(4):319–28. DOI: https://doi.org/10.1001/jama.2024.21700.

12. Adam GP, DeYoung J, Paul A, Saldanha IJ, Balk EM, Trikalinos TA, Wallace BC. Literature search sandbox: a large language model that generates search queries for systematic reviews. JAMIA Open. 2024;7(3):ooae098. DOI: https://doi.org/10.1093/jamiaopen/ooae098.

13. Bourgeois JP, Ellingson H. Ability of ChatGPT to generate systematic review search strategies compared to a published search strategy. Med Ref Serv Q. 2025 Jul-Sep:44(3):279-291. DOI: https://doi.org/10.1080/02763869.2025.2537075.

14. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval [Internet]. Taipei, Taiwan: ACM; 2023 [cited 23 Sept 2025]. p. 1426–36. https://dl.acm.org/doi/10.1145/3539618.3591703.

15. Akinseloyin O, Jiang X, Palade V. A question-answering framework for automated abstract screening using large language models. J Am Med Inform Assoc. 2024 Sept 1;31(9):1939-1952. DOI: https://doi.org/10.1093/jamia/ocae166.

16. Lieberum JL, Toews M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, Böhringer D, Meerpohl JJ, Eisele-Metzger A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol. 2025 May;181:111746. DOI: https://doi.org/10.1016/j.jclinepi.2025.111746.

17. Department of Biomedical Informatics Generative AI at VUMC [Internet]. Vanderbilt University Medical Center; [cited 23 Sept 2025]. https://www.vumc.org/dbmi/GenerativeAI.

18. Blasingame MN, Koonce TY, Williams AM, Giuse DA, Su J, Krump PA, Giuse NB. Evaluating a large language model's ability to answer clinicians' requests for evidence summaries. J Med Libr Assoc. 2025 Jan 14;113(1):65–77. DOI: https://doi.org/10.5195/jmla.2025.1985.

19. Koonce TY, Williams AM, Giuse DA, Su J, Blasingame MN, Krump PA, Giuse NB. A multi-model evaluation: harnessing generative AI to understand the state-of-the-art of literature search automation. Medical Library Association Annual Meeting, Pittsburgh, PA; Apr 2025.

20. The CHART Collaborative; Huo B, Collins GS, Chartash D, Thirunavukarasu AJ, Flanagin A, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open. 2025 Aug 1;8(8):e2530220. DOI: https://doi.org/10.1001/jamanetworkopen.2025.30220.

21. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ. 2025 Aug 1;390:e083305. DOI: https://doi.org/10.1136/bmj-2024-083305.

22. Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. Int J Med Inf. 2024 Aug;188:105474. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105474.

23. OpenAI. ChatGPT free tier FAQ [Internet]. OpenAI Help Center; [cited 2025 Aug 1]. https://help.openai.com/en/articles/9275245-chatgpt-free-tier-faq.

24. Kavukcuoglu K. Gemini 2.0 is now available to everyone [Internet]. Google; 5 Feb 2025 [cited 23 Sept 2025]. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/.

25. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948 [Preprint]. arXiv; 2025 [cited 23 Sept 2025]. Available from: http://arxiv.org/abs/2501.12948.

26. xAI. Grok 3 Beta — the age of reasoning agents [Internet]. xAI; 19 Feb 2025 [cited 23 Sept 2025]. https://x.ai/news/grok-3.

27. Spataro J. Available today: GPT-5 in Microsoft 365 Copilot [Internet]. Microsoft; 7 Aug 2025 [cited 17 Oct 2025]. https://www.microsoft.com/en-us/microsoft-365/blog/2025/08/07/available-today-gpt-5-in-microsoft-365-copilot/.

28. GovTech Data Science & AI Division. Prompt engineering playbook (Beta v3) [Internet]. Government of Singapore; 30 Aug 2023 [cited 23 Sept 2025]. https://www.developer.tech.gov.sg/products/collections/data-science-and-artificial-intelligence/playbooks/prompt-engineering-playbook-beta-v3.pdf.

29. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007 June 15;7(1):16. DOI: https://doi.org/10.1186/1472-6947-7-16.

30. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12-13.

31. American Diabetes Association Professional Practice Committee for Diabetes. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2026. Diabetes Care. 2026 Jan 1;49(Suppl 1):S27-S49. DOI: https://doi.org/10.2337/dc26-S002.

32. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377–81. DOI: https://doi.org/10.1016/j.jbi.2008.08.010.

33. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN, REDCap Consortium. The REDCap consortium: building an international community of software platform

DOI: dx.doi.org/10.5195/jmla.2026.2333

partners. J Biomed Inform. 2019 July;95:103208. DOI: https://doi.org/10.1016/j.jbi.2019.103208.

34. Tao D, Kochendorfer KM, Griffin T, McCrary Q, Gautam A, Labib BSR, Arvan M, Flynn J, Jiang K. "Can ChatGPT answer patient's questions?": a preliminary analysis. Stud Health Technol Inform. 2025 Aug 7;329:1586–7. DOI: https://doi.org/10.3233/shti251114.

35. Fox ZE, Williams AM, Blasingame MN, Koonce TY, Kusnoor SV, Su J, Lee P, Epelbaum MI, Naylor HM, DesAutels SJ, Frakes ET, Giuse NB. Why equating all evidence searches to systematic reviews defies their role in information seeking. J Med Libr Assoc. 2019 Oct 1;107(4):613–7. DOI: https://doi.org/10.5195/jmla.2019.707.

36. Lin CR, Chen YJ, Tsai PA, Hsieh WY, Tsai SHL, Fu TS, Lai PL, Chen JY. Multiple large language models versus clinical guidelines for postmenopausal osteoporosis: a comparative study of ChatGPT-3.5, ChatGPT-4.0, ChatGPT-4o, Google Gemini, Google Gemini Advanced, and Microsoft Copilot. Arch Osteoporos. 2025 Sept 8;20(1):120. DOI: https://doi.org/10.1007/s11657-025-01587-4

37. Flaharty KA, Hu P, Hanchard SL, Ripper ME, Duong D, Waikel RL, Solomon BD. Evaluating large language models on medical, lay-language, and self-reported descriptions of genetic conditions. Am J Hum Genet. 2024 Sept 5;111(9):1819–33. DOI: https://doi.org/10.1016/j.ajhg.2024.07.011.

38. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects. J Am Med Inform Assoc. 2002;9(1):1–15. DOI: https://doi.org/10.1136/jamia.2002.0090001.

39. Bockting CL, van Dis EAM, van Rooij R, Zuidema W, Bollen J. Living guidelines for generative AI - why scientists must oversee its use. Nature. 2023 Oct;622(7984):693–6. DOI: https://doi.org/10.1038/d41586-023-03266-1.

40. Cardero R, Sarro E. AI fluency as an essential element towards a smarter workforce [Internet]. HRD; 30 Aug 2025 [cited 23 Sept 2025]. https://www.hrdconnect.com/2025/08/30/ai-fluency-as-an-essential-element-towards-a-smarter-workforce/.

41. Robinson K, Bontekoe K, Muellenbach J. Integrating PICO principles into generative artificial intelligence prompt engineering to enhance information retrieval for medical librarians. J Med Libr Assoc. 2025 Apr 18;113(2):184–8. DOI: https://doi.org/10.5195/jmla.2025.2022.

42. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc AMIA Symp. 2006;2006:359–63.

43. Stanford Center for Research on Foundation Models. MedHELM - Holistic Evaluation of Language Models (HELM) [Internet]. Stanford University; 2 Jun 2025 [cited 2025 Sept 23]. https://crfm.stanford.edu/helm/medhelm/latest/#/leaderboard.

44. Bean AM, Payne R, Parsons G, Kirk HR, Ciro J, Mosquera R, Monsalve SH, Ekanayaka AS, Tarassenko L, Rocher L, Mahdi A. Clinical knowledge in LLMs does not translate to human interactions. arXiv:2504.18919 [Preprint]. arXiv;2025 Apr [cited 24 Sept 2025]. Available from: http://arxiv.org/abs/2504.18919.

45. Rebitschek FG, Carella A, Kohlrausch-Pazin S, Zitzmann M, Steckelberg A, Wilhelm C. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. NPJ Digit Med. 2025 June 9;8(1):343. DOI: https://doi.org/10.1038/s41746-025-01752-6.

SUPPLEMENTAL FILES

Appendix A: CHART Checklist and Methodological Diagram
Appendix B: List of PICO Questions
Appendix C: Prompt for PICO Questions Generation
Appendix D: Prompt for Submitting the Questions to the LLMs
Appendix E: Prompt for Obtaining the Key Elements

AUTHORS' AFFILIATIONS

Mallory N. Blasingame, mallory.n.blasingame@vumc.org, https://orcid.org/0000-0003-0356-9481, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Taneya Y. Koonce, taneya.koonce@vumc.org, https://orcid.org/0000-0002-4014-467X, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Annette M. Williams, annette.williams@vumc.org, https://orcid.org/0000-0002-2526-3857, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Jing Su, jing.su@vumc.org, https://orcid.org/0000-0001-6699-6806, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Dario A. Giuse, dario.giuse@vumc.org, https://orcid.org/0000-0002-2677-6734, Department of Biomedical Informatics, Vanderbilt University School of Medicine, Vanderbilt University Medical Center, Nashville, TN

Poppy A. Krump, poppy.krump@vumc.org, https://orcid.org/0000-0002-3081-6487, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Nunzia B. Giuse, nunzia.giuse@vumc.org, https://orcid.org/0000-0002-7644-9803, Center for Knowledge Management, Department of Biomedical Informatics, and Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Received September 2025; accepted December 2025

ISSN 1558-9439 (Online)