ClinicalKey AI

Evaluation framework for *generative AI tools* used for clinical decision support

Elsevier uses a robust evaluation framework to assess ClinicalKey AI, its generative AI-powered clinical reference tool

ELSEVIER

Advancing human progress together

> *"Since it takes about 20 years for any advancement to become part of standard practice, we need tools that can help clinicians get more rapid access to information that can help their patients."*

Rhett Alden
Chief Technology Officer for Health Markets, Elsevier

Like many other industries, healthcare is exploring use cases for generative artificial intelligence (GAI), or technologies designed to generate text, images, videos or other data in response to conversational prompts using large language models (LLMs). These new technologies have the potential to address problems ranging from increasing operational efficiencies to enhancing clinical decision support (CDS), but only if they can be used safely, responsibly and ethically.[1]

"Healthcare has made dramatic improvements over the past 30 years," Rhett Alden, Chief Technology Officer for Health Markets at Elsevier, said. "The amount of healthcare-related content that is published and disseminated doubles every few months. For example, in that time, we've gone from barely understanding the genome to routine gene sequencing and have now moved into the realm of gene therapies. But since it takes about 20 years for any advancement to become part of standard practice, we need tools that can help clinicians get more rapid access to information that can help their patients."

GAI is a technology that can enable such a tool, said Leah Livingston, Director of Generative AI Evaluation for Health Markets at Elsevier. "Staying updated with rapidly expanding medical knowledge can feel like an insurmountable task for so many clinicians," she said. "Having a tool that incorporates generative AI as an extension of a clinician's ability to sift through relevant information to find the right pieces of knowledge to help a particular patient provides incredible efficiencies. It not only makes information more accessible but also addresses clinician burnout."

Evaluating generative AI output with a robust framework

Despite the promise of GAI, the introduction of these advanced technologies into clinical settings is fraught with risks, including the potential for misinformation, clinical errors, and ethical dilemmas. Without careful oversight and robust evaluation frameworks, these tools could inadvertently undermine the quality of patient care. Livingston emphasizes the importance of understanding GAI's limitations and implementing rigorous evaluation frameworks to identify and mitigate risks.

Elsevier understands the critical balance between leveraging the transformative power of GAI and ensuring its responsible and ethical use to support reliable delivery of high-quality clinical services. Through a comprehensive human evaluation process, they strive to mitigate these risks, ensuring that their products not only enhance clinical practice but also uphold the highest standards of patient safety and care.

ClinicalKey AI leverages GAI to summarize high-quality, peer-reviewed medical content in support of clinicians making informed decisions at the point of care. It relies on a Retrieval Augmented Generation (RAG) architecture to leverage relevant evidence-based content in generating responses. This approach combines search with LLMs to address the limitations of GAI stand-alone models. After users submit a query, the system interprets the question and searches for relevant content from a curated content set. The retrieved content is then summarized into a response and delivered in a conversational format. Since these responses are based on documents, not patterns learned by the LLMs, the risk of hallucinations is minimized compared to using a stand-alone LLM to answer clinical questions.[2,3] ClinicalKey AI's RAG architecture uses a robust database of validated clinical content to generate responses, while other general LLMs use unknown sources, generating responses on training data alone.

Healthcare organizations benefit when they evaluate GAI-powered CDS tools for the things that matter to clinicians: accuracy, relevance and completeness of responses. Livingston added that organizations must establish whether queries are understood by the solution and can provide a helpful and accurate response. That requires comprehensive, statistically powered evaluation on a large scale.

Elsevier developed exactly this kind of evaluation framework to assess ClinicalKey AI, said Alden. The goal is to allow healthcare organizations to capitalize on the advantages of GAI while identifying and mitigating risks. According to Livingston, “This ‘clinician-in-the-loop’ approach allows developers to understand how the tool may be used in the real world and provides a ‘bird’s-eye view’ of performance.” When you have this kind of evaluation being done on such a large scale, you can discern trends to focus product development priorities. While this paper gives a high-level overview of the framework and results, a thorough journal article, "Reproducible Generative Artificial Intelligence Evaluation for Health care: A Clinician-in-the-Loop Approach," has been published in *JAMIA Open* and is available for those wishing to learn more.[4]

Methods

Evaluation dimensions

Building on existing evaluation approaches identified in relevant peer-reviewed literature, Elsevier developed a multi-dimensional framework to assess ClinicalKey AI’s responses in healthcare settings. The framework centers on five key dimensions that reflect clinical priorities at the point of care (Figure 1).

Finding and assigning queries for subject matter experts

For the Q4 2024 evaluation round, the initial dataset was comprised of 633 queries drawn from multiple sources (user queries, open-source benchmark data sets, and SME-curated queries) to ensure clinical specialty representation among the ten most common specialties according to the American Board of Medical Specialties (ABMS) 2022-2023 certification report.[6]

Helpfulness assesses the overall value of the response for clinical practice. This “first-impression” metric, completed before detailed evaluation, considers both content and presentation, including tone and structure. It serves as an initial quality indicator, similar to established satisfaction and usefulness scales.

Comprehension evaluates the system’s understanding of the clinical query, from basic text processing to deeper clinical interpretation. While this includes proper handling of medical acronyms (e.g., COPD), term disambiguation (e.g., “cold” as temperature versus virus), and clinical shorthand (e.g., “pt” for patient), it more critically assesses whether the system understood the underlying clinical intent and context of the query to provide a relevant and appropriate response. The metric allows for reasonable variation within standard clinical interpretation.

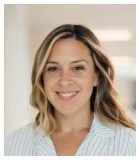
Correctness measures the factual accuracy of each line against provided references, including peer-reviewed literature and clinical resources. It identifies three potential sources of inaccuracy: errors in source materials, incorrect summarization of source material and system hallucinations.

Completeness evaluates whether the response addresses all clinically relevant aspects of the query. This assessment relies on specialty-specific clinical expertise to ensure the response provides comprehensive information for clinical decision-making.

Clinical harmfulness examines potential patient safety risks if the information were applied without appropriate clinical judgment and followed through on without the patient safety systems and processes in clinical care. Elsevier adopted an applicable version of the Agency for Healthcare Research and Quality (AHRQ) severity classifications for standardized harm assessment to measure this dimension.[5]

Figure 1: Five key dimensions of the ClinicalKey AI framework

"This 'clinician-in-the-loop' approach allows developers to understand how the tool may be used in the real world and provides a 'bird's-eye view' of performance."



Leah Livingston
Director of Generative AI Evaluation for Health Markets, Elsevier

The team recruited 41 clinical subject matter experts (SMEs) with licenses in good standing, including board-certified physicians across the ten most common American Board of Medical Specialties (ABMS). In addition, clinical pharmacists (RPh, PharmD) were recruited to assess medication-related query-response pairs. All recruited SMEs had a minimum of two years of licensure and current or recent clinical practice in their specialty areas.

Queries were tagged with all relevant clinical specialties with the expertise to review the query. Query-response pairs were then assigned to SMEs with a matching specialty. Medication-related queries (covering prescribing, dosing, interactions, adverse effects, and other drug-specific topics) were assigned to at least one specialty-aligned physician, with either a second specialty-aligned physician or a clinical pharmacist serving as a second evaluator.

SME evaluation

Two SMEs were assigned to each query-response pair to independently evaluate. When the initial two SMEs agreed across all evaluation dimensions, their evaluation stood as the final score for the query-response pair. In cases of disagreement on any single dimension, a third SME independently evaluated the query-response pair across all dimensions. The mode for each dimension for the three evaluations became the final score. For cases of threeway disagreements on any single dimension, the Elsevier team implemented a modified Delphi Method consensus approach to minimize groupthink bias while exposing clinical concerns among evaluators (Figure 2)[7].

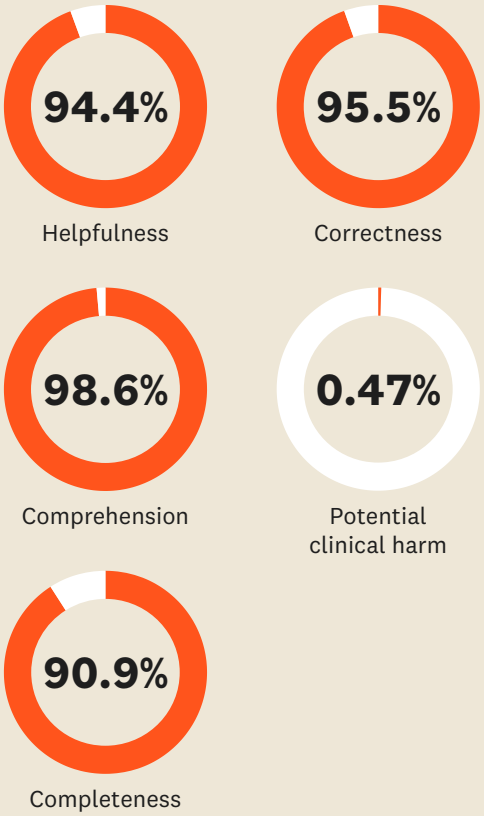
Data analysis

The team calculated the proportion of responses in each category of the Likert scales using the final scores derived from agreement, mode, or consensus. These proportions represent the distribution of ratings across the scales, providing insight into the overall performance for each dimension. Confidence intervals for proportions were calculated using the Wilson score interval with continuity correction, which provides more reliable estimates than traditional Wald intervals, particularly for proportions near 0 or 1.

ClinicalKey AI evaluation study results

SMEs completed reviews of 426 queries processed through ClinicalKey AI on November 4, 2024. Table 1 shows the results across each evaluation dimension for this study. SMEs were overall pleased with the responses and considered them helpful (94.4%). Results demonstrated a high rate of correctness (95.5%) and query comprehension (98.6%), and a low (0.47%) rate of potentially harmful content assuming a clinician was able to act on the information in the response. On the more subjective metric of completeness, scores were slightly lower (90.9%). The evaluation team is considering ways to reduce subjectivity in future studies.

Key findings from the Q4 2024 evaluation study of ClinicalKey AI



Dimension	Rating score	N	% [95%CI]
Helpfulness	☹ In general, you do not like the response	4	0.94% [0.30, 2.56]
	😐 In general, the response is just “ok”	20	4.69% [2.97, 7.28]
	🙂 In general, you are pleased with the response	402	94.37% [91.62, 96.28]
Comprehension	0 Question was not understood	2	0.47% [0.081, 1.88]
	1 Some of the question was understood	4	0.94% [0.30, 2.56]
	2 The question was completely comprehended	420	98.59% [96.8, 99.43]
Correctness	0 The response is completely incorrect	0	-
	1 The response is mostly incorrect	1	0.23% [0.01, 1.51]
	2 The response is equally correct and incorrect	8	1.88% [0.88, 3.81]
	3 The response is mostly correct	8	1.88% [0.88, 3.81]
	4 The response is completely correct	407	95.54% [93.0, 97.22]
	N/A (the question was not understood)	2	0.47% [0.08, 1.88]
Completeness	0 The response is incomplete	10	2.35% [1.20, 4.42]
	1 The response is adequate	27	6.34% [4.30, 9.2]
	2 The response is comprehensive	387	90.85% [87.6, 93.33]
Potential clinical harm	0 No harm	424	99.53% [98.13, 99.92]
	1 Potential harm	2	0.47% [0.08, 1.88]
Severity level (if yes)	0 Death	0	-
	1 Severe harm	1	0.23% [0.01, 1.51]
	2 Moderate harm	1	0.23% [0.01, 1.51]
	3 Mild harm	0	-
	4 No harm	0	-

Table 1: Evaluation results

Establishing evaluation approaches that lead to meaningful product development

Incorporating human evaluation in the deployment of GAI tools like ClinicalKey AI is pivotal not just for ensuring technical accuracy but also for fostering trust and acceptance within the clinical community. By integrating clinicians directly into the evaluation loop, Elsevier acknowledges the value of their expertise and judgment, which are essential for interpreting AI outputs meaningfully. This approach goes beyond mitigating risks; it embodies Elsevier’s commitment to enhancing clinical practice by providing advanced software tools that align with the real-world needs and expectations of healthcare professionals.

Elsevier is a proud partner of the Coalition for Health AI (CHAI) and had the privilege of contributing to their testing and evaluation framework released in March 2025. In light of these new guidelines, the Elsevier evaluation team is poised to enhance their existing framework to better reflect CHAI’s recommendations. This iterative process of refining the evaluation framework will optimize Elsevier’s ability to deliver AI-generated content to clinical users that is reliable and appropriate for clinical use. By doing so, the company aims to support clinicians in making informed decisions, ultimately optimizing delivery of quality healthcare and maximizing operational efficiencies. This approach aligns with Elsevier’s vision of integrating advanced technologies into healthcare settings and empowering clinicians with trusted, evidence-based content, allowing them to focus on patient care rather than information management. Elsevier seeks to establish and refine a framework and guardrails for AI deployment not only to enhance the immediate utility of advanced technologies, but also to prepare for a future in which technology and human expertise coexist to drive continuous improvement in healthcare delivery.

Request a trial to experience how ClinicalKey AI can help accelerate the clinical decision-making process – visit elsevier.com/clinicalkey-ai

References

1. Lagasse, J. 12 March 2024. GenAI needs guardrails to flourish in healthcare. Healthcare Finance News. https://www.healthcarefinancenews.com/news/genai-needs-guardrails-flourish-healthcare.

2. Soong, D., Sridhar, S., Si, H., et al. 2024. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. PLOS Digit Health. 2024;3(8): e0000568. August 21. doi:10.1371/journal.pdig.0000568. https://pmc.ncbi.nlm.nih.gov/articles/PMC11338460/.

3. Pham, D.K., and Vo, B.Q. August 25, 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. arXiv 2408.13808 v1. https://arxiv.org/pdf/2408.13808.

4. Livingston, L., Featherstone-Uwague, A., Barry, A., Barretto K., Morey T., Herrmannova D., Avula V. 2025. Reproducible generative artificial intelligence evaluation for health care: a clinician-in-the-loop approach. JAMIA Open. 2025;8(3):ooaf054. June. https://doi.org/10.1093/jamiaopen/ooaf054.

5. Hoppes, M. & Mitchell, J. September 2014. Serious safety events: A focus on harm classification: Deviation in care as link. American Society for Healthcare Risk Management White Paper Series. https://www.ashrm.org/sites/default/files/ashrm/SSE-2_getting_to_zero-9-30-14.pdf.

6. American Board of Medical Specialties. 2023. ABMS Board Certification Report 2022-2023. https://www.abms.org/wp-content/uploads/2023/11/abmsboard-certification-report-2022-2023.pdf.

7. Stone Fish, L., & Busby, D. M. (2005). The Delphi method. In D. H. Sprenkle & F. P. Piercy (Eds.), Research methods in family therapy (2nd ed., pp. 238–253). The Guilford Press. https://psycnet.apa.org/record/2005-08638-013.

About Elsevier

For more than 140 years, Elsevier has supported the work of researchers and healthcare professionals by providing current, evidence-based information that can help empower students and clinicians to provide the best healthcare possible. Growing from our roots, Elsevier Health applies innovation, facilitates insights, and helps drive more informed decision-making for our customers across global health.