ClinicalKey AI

Helping to mitigate risk in generative AI tools for clinical decision support

Elsevier uses a robust evaluation framework to assess ClinicalKey AI, its generative AI-powered clinical decision support tool, for efficacy and bias

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]

Rhett Alden, Chief Technology Officer for Health Markets at Elsevier, said such offerings have been at the forefront of technology conversations for the past few years because of their promise to help manage the "sheer explosion of information," particularly in the healthcare space.

"Healthcare has made dramatic improvements over the past 30 years," he 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, Senior Clinical Data Scientist and Responsible AI Expert at Elsevier. And with the high rates of burnout that so many clinicians are experiencing, a reliable tool enabling them to easily stay current with relevant medical information could benefit all healthcare stakeholders. Such a tool could help provide higher quality care to patients, lighten workloads for clinicians and increase overall value and efficiency for healthcare organizations.

"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 research and 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 to enhance patient outcomes but also addresses the problem all of healthcare is having with clinicians being overwhelmed."

> *"For a high-stakes environment like healthcare, we need to be able to minimize the risk of inaccurate responses being returned to clinicians."*

Leah Livingston

Senior Clinical Data
Scientist and Responsible
AI Expert, Elsevier

Evaluating generative AI output with a robust framework

Despite the promise of GAI, helping to ensure its safe, responsible, and ethical use in clinical settings remains a challenge. Livingston emphasizes the importance of understanding GAI's limitations and implementing rigorous evaluation frameworks to help mitigate risks and protect patients.

"We know that these models come with a risk of something known as 'hallucinations,' which is a technology industry term for inaccurate content being represented in a response," said Livingston. "For a high-stakes environment like healthcare, we need to be able to minimize the risk of inaccurate responses being returned to clinicians."

ClinicalKey AI, a CDS tool that leverages GAI in summarizing high-quality, peer-reviewed medical content to support clinicians in making informed decisions at the point of care, relies on a Retrieval Augmented Generation (RAG) architecture to manage this issue. This approach, which is quickly becoming accepted in the industry, combines search with LLMs to address the limitations of GAI models. After users 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 minimal. In contrast, general LLMs generate responses in a probabilistic or variable way based on patterns in the text it was trained on; there is no database of facts behind the response.

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 in a clinically appropriate way. That requires comprehensive, statistically powered evaluation on a large scale.

Elsevier developed exactly this kind of evaluation framework to assess and validate its latest solution, ClinicalKey AI, said Alden (Figure 1). The goal is to allow healthcare organizations to capitalize on the advantages of GAI while managing risks.

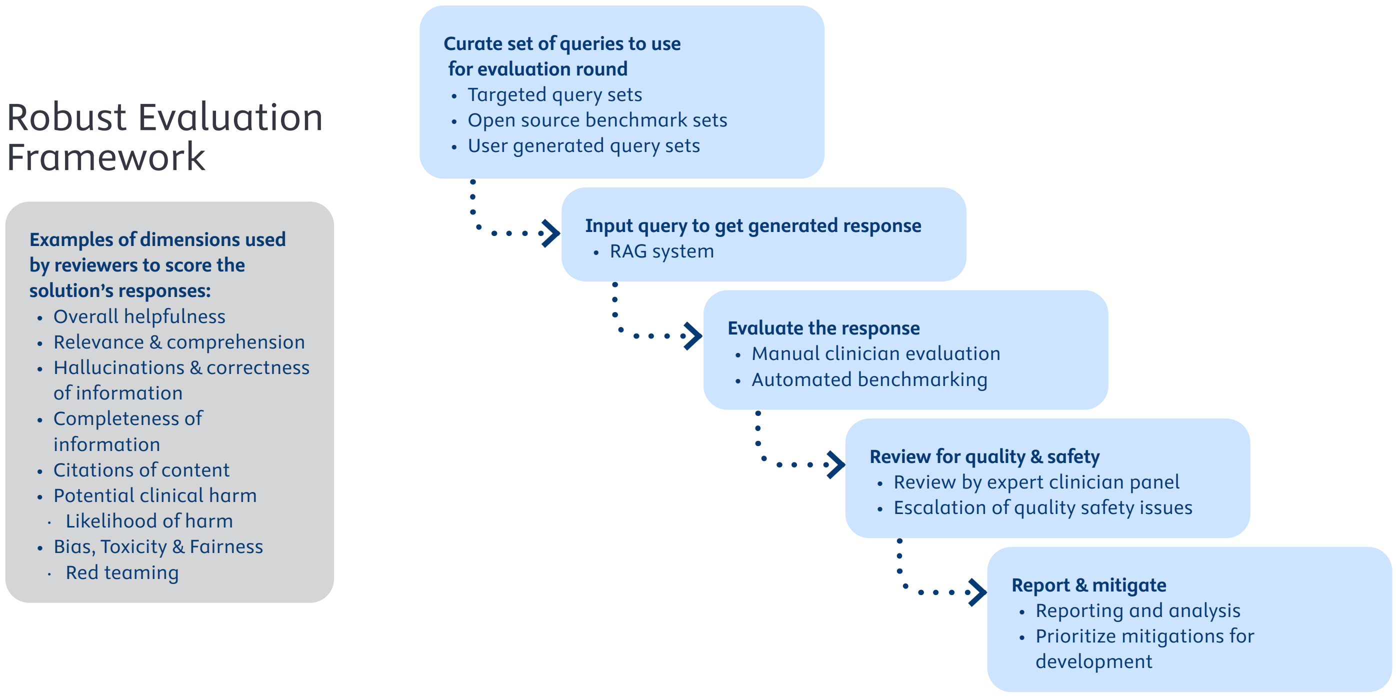
How Elsevier evaluates ClinicalKey AI responses for quality

Elsevier's evaluation framework for ClinicalKey AI is a cornerstone of its development, helping to ensure the tool meets the high standards required for clinical use. Evaluations are performed quarterly and before every major release. The framework involves several key steps.

1. **Multidimensional scoring by clinicians:** The box at the left of Figure 1 contains some examples of dimensions for clinicians' scoring of the solution's responses. Clinicians from a variety of specialties rate each response on a range of different dimensions, including the correctness and completeness of the information presented, and determine whether following a potentially flawed response would carry potential for harm. This controlled study is crucial for refining the tool.

2. **Manual review and expert analysis:** The diagram to the right illustrates the process of evaluating ClinicalKey AI's responses. About 2,000 queries are obtained from practicing clinicians and from open-source datasets. Queries are processed by the RAG system (referenced above in the paper). Each response is then manually evaluated by clinicians. Any responses on which there is significant disagreement are sent to a panel of expert clinicians for further analysis.

3. **Continuous improvement:** The full set of evaluated responses is analyzed and used to quantify the performance of the system as a whole. Any issues identified are used to guide further improvements. Feedback from clinicians is also used to continuously improve the tool, helping to ensure it evolves to meet the changing needs of the healthcare environment.



Robust Evaluation Framework

Examples of dimensions used by reviewers to score the solution's responses:

- Overall helpfulness
- Relevance & comprehension
- Hallucinations & correctness of information
- Completeness of information
- Citations of content
- Potential clinical harm
  - Likelihood of harm
- Bias, Toxicity & Fairness
  - Red teaming

Curate set of queries to use for evaluation round

- Targeted query sets
- Open source benchmark sets
- User generated query sets

Input query to get generated response

- RAG system

Evaluate the response

- Manual clinician evaluation
- Automated benchmarking

Review for quality & safety

- Review by expert clinician panel
- Escalation of quality safety issues

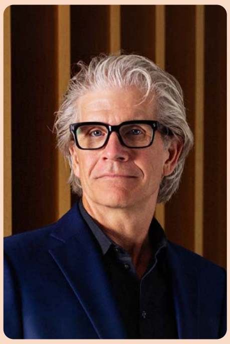
Report & mitigate

- Reporting and analysis
- Prioritize mitigations for development

Figure 1. The framework Elsevier uses to evaluate ClinicalKey AI, its generative AI-powered CDS tool.

ELSEVIER

ClinicalKey AI



“We are leaning in on trust and AI ethics so we can reduce risk and bias to ensure that clinicians are getting the value they need from the tool.”

Rhett Alden

Chief Technology Officer
for Health Markets,
Elsevier

The evaluation team recruited more than one hundred licensed clinicians across major medical specialties to review query-directed responses from the solution. Subject matter experts underwent training on guidelines for GAI, ClinicalKey AI’s system and the evaluation framework. The evaluation team was provided with practice rounds to assess their understanding of feedback methods before they participated in an evaluation round. This surveillance activity provided insights into the performance of ClinicalKey AI, allowing the team to determine any refinements that needed to be made. Alden added that the team also applied a risk management framework commonly used for medical devices to the solution, even though the solution is not classified as one.

“We are approaching this from a probability of harm perspective,” he said. “That’s why we put mitigations and controls in place to help ensure that risks are reduced.”

According to Livingston, this “human 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” so the team can be confident of its performance before putting it in users’ hands. “When you have this kind of evaluation being done on such a large scale, you can discern why something might not be working the way it should be. We can dig into the details and see why things are happening the way they are happening, and then we can work to improve the tool,” she said.

Addressing potential bias and building trust

Elsevier’s comprehensive evaluation framework also allows the organization to assess ClinicalKey AI for potential bias. Alden said that the susceptibility of GAI to bias was a hot topic at the recent 2024 HIMSS Global Health Conference & Exhibition in Orlando, Florida. When models are trained on data sets that do not represent diverse patient populations, it can lead to answers to care questions that may not benefit, for example, minority populations. Bias in GAI output can be more nuanced and complex to identify and mitigate. Alden and his team are making sure to include bias in their evaluation criteria, constructing a bias framework to deal with issues that may affect clinical decisions involving underrepresented groups. The evaluation team also adheres to Elsevier’s responsible AI principles to help make sure that ethical concerns are being addressed throughout development. Elsevier closely monitors the AI landscape in healthcare to ensure it understands regulatory obligations, industry expectations and customers' needs.

"By putting this all together, our goal is to create a tool that any clinician can point to and say, 'there's transparency here,'" he said. "We are leaning in on trust and AI ethics so we can reduce risk and bias to ensure that clinicians are getting the value they need from the tool."

Livingston added that she and her colleagues worked hard to create a methodical evaluation framework with the objective of assessing the value for clinicians. "As a former clinician, I know the questions I would be asking before I would comfortably use a tool like this myself," she said. "The goal was also to build a framework that could answer those questions to make clinicians confident in understanding the risks, as well as the benefits, of using a tool like this."

Putting proper governance in place

GAI is here to stay, Alden said. "Generative AI technologies are developing quickly. Hospitals are under tremendous pressure to determine a governance structure to help clinicians understand how, where and when to use them — but it can be a daunting task," he added. "There isn't one right way to do this. But some clinicians are already using these tools without such policies in place. It's important for healthcare organizations to start moving forward with evaluation and governance now so they can make more informed decisions for their institutions moving forward."

Livingston agreed and added that, as ClinicalKey AI is an evolving entity with new content being added daily, she and her team will continue to evaluate it using the robust framework to test accuracy and reliability with attention towards reducing risk and bias.

"This work is built into the product development framework we have going forward," she explained. "We are consciously aware that any differences can impact the end user. It's important to us that we are providing trustworthy versions of generative AI tools so we can put clinicians in a better position to provide more efficient and thorough care to patients moving forward."

Reference

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.

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.