

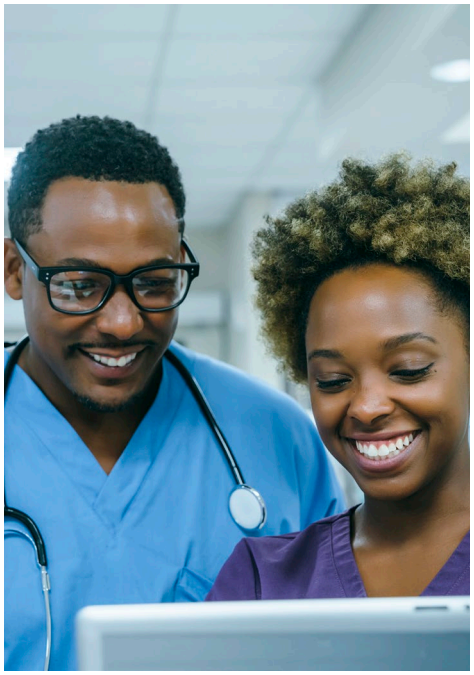
A Measured Approach to Evaluating AI at the Point of Care



How UpToDate Expert AI is assessed for clinical usefulness, knowledge integrity, and potential risks

Todd Dardas, MD, Adam Rabinowitz, Evgeniy Padaliak, Sheila Bond, MD

Wolters Kluwer



Clinicians should know how trustworthy their AI is.

UpToDate® has served clinicians for more than 30 years as a trusted point-of-care reference. **UpToDate**® **Expert AI**, our generative AI solution, extends that foundational clinical decision support (CDS) by allowing clinicians to interact with UpToDate's curated clinical knowledge in a new way: through more flexible queries, more tailored responses, and faster access to relevant information.

Evaluation of AI-generated clinical content must go beyond technical performance. Because the words our system produces may influence clinical decisions, we determined that available industry benchmarks, user ratings, and generic scores were not enough to be reliable. We built a custom, multi-method evaluation approach to directly assess the words clinicians actually read.

UpToDate Expert AI is a clinical decision support tool in the United States and certain other jurisdictions. It is a reference-only tool in others.

The dimensions we measure

UpToDate Expert AI answers are directly evaluated across three distinct core dimensions:

- **Clinical intent:** Is the answer faithful to our clinical content and point-of-care standards?
- **Knowledge integrity:** Is the answer grounded in trusted clinical knowledge?
- **Potential risks**: How does the system behave under stress, uncertainty, and complex clinical scenarios?

Together, these dimensions provide a more complete assessment of reliability than any single benchmark or score, and help to evaluate the answer itself: what it says, what it leaves out, what it adds, and how it behaves under stress or in real clinical settings (Figure 1).



Figure 1: *Three core dimensions that we measure*

Clinical intent: Is the answer faithful to our clinical content and point-of-care standards?

Rubric testing:

We evaluate UpToDate Expert AI answers against physician-authored rubrics that define the elements of an ideal answer for a given query. These rubrics are created and maintained by UpToDate physician editors and span 25 medical specialties, all venues of care, and the spectrum of query complexity (Figure 2).

Rubric-based measurement

At UpToDate, we test the performance of UpToDate Expert AI with clinical rubrics composed of **1,600 queries** with more than **15,000 criteria** from across **25 medical specialties.**

Rubrics are physician-authored tools created specifically to evaluate AI-generated answers for point-of-care use. Each rubric is composed of a clinical query and the criteria that define what should be included in a good answer.

Rubrics allow us to break the open-ended answers from UpToDate Expert AI into specific, independently scored elements that assess whether the answer:

- Includes expected clinical information
- Has relevant clinical context
- Contains incorrect information
- Omits details important for point-of-care use

This provides us with detailed information on what is present, what is missing, and where an answer may be incomplete or unsupported.

UpToDate Expert AI provided clinically aligned information for 99.9% of assessed criteria

Mock rubric

Question: *"When to anticoagulate for atrial fibrillation?"*

Criteria What should the answer include for point-of-care use?	**Importance** How important is this component to the answer?	**Impact** What is the impact if incorrect or missed?
Describes the CHA_2DS_2-VASc thresholds for males and females	Tier 1	Higher
Must mention an assessment of bleeding risk	Tier 1	Higher
Must mention that all patients are anticoagulated after DC cardioversion	Tier 1	Higher
Anticoagulation is mandatory for 4-6 weeks after DC cardioversion	Tier 2	Lower
The $CHADS_2$ score is no longer used	Tier 3	Lower

Figure 2: *Schematic of the clinical rubric*

In the most recent evaluation, UpToDate Expert AI was tested on 1,669 clinical queries comprising more than 15,000 criteria. **UpToDate Expert AI provided clinically aligned information for 99.9% of assessed criteria**. Rubric testing also allows us to detect omissions: clinically meaningful information expected under the rubric that did not appear in the answer. UpToDate Expert AI had a significantly lower rate of omissions when compared with two general-purpose LLM comparators; both comparators had a rate of omission that was 15% higher than UpToDate Expert AI ($p<0.0001$ for both comparisons to UpToDate Expert AI). This difference translates to users seeing one additional error for every seven queries when using a general-purpose LLM model when compared with the use of UpToDate Expert AI.

UpToDate Expert AI met more Essential criteria than general-purpose LLM comparators, with 13%–15% higher rates of Essential criteria met. This pattern has been observed across multiple evaluations during the development of the UpToDate Expert AI system.

In addition to supporting iterative development, this test is also used to monitor for regression or drift during iterative improvements to the system.



Knowledge integrity: Is the answer grounded in valid, trusted clinical knowledge?

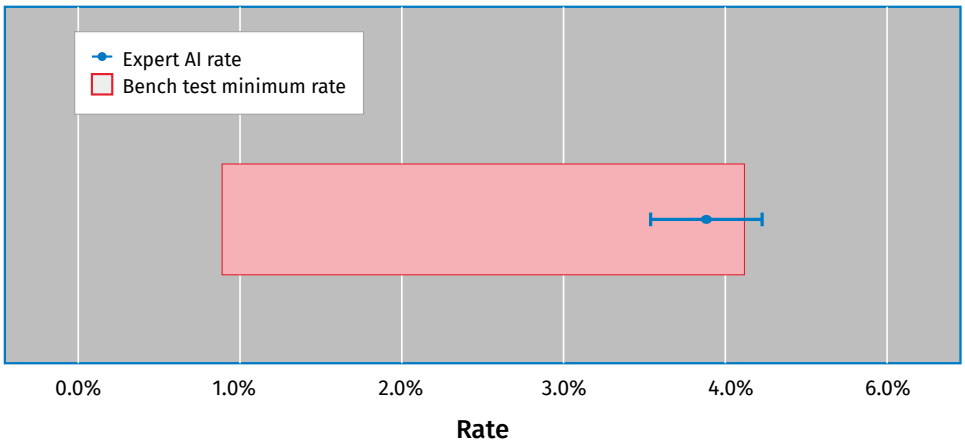
UpToDate Expert AI is intended to produce answers grounded in UpToDate content. The system prompt specifies that only UpToDate content can be used, employs a multi-step retrieval process, and is intended to avoid responding if relevant UpToDate content cannot be located. Despite this design, the same generative technology that enables more natural queries, interpretation of abbreviations, state-of-the-art search capability, and answer synthesis can introduce information from outside of UpToDate into answers, which we refer to as model knowledge. Since model knowledge can sound plausible and is sometimes even correct (e.g., the translation of a non-standard abbreviation), detecting model knowledge can be difficult.

Our approach to testing knowledge integrity is based on a custom test that assesses whether the words in an answer were derived directly from UpToDate's trusted clinical content. In a recent estimate of model knowledge use that included 1066 queries from across the breadth of medicine, the rate of model knowledge was 3.9% (608 possible model knowledge claims among 15,681 claims evaluated), which is similar to our test's background noise rate of 1-4% as determined from bench testing. This estimate conservatively includes some model knowledge that is factually correct. This value suggests that **UpToDate Expert AI is focused on trusted content within UpToDate** and that our system operates true to its purpose.



UpToDate Expert AI is grounded in trusted content within UpToDate

Model Knowledge Rates



Note: Model knowledge rates include correct and incorrect information as well as unavoidable false positive claims.

Figure 3: *Model knowledge and background noise*

Potential risks: How does the system behave under stress, uncertainty, and complex clinical scenarios?

We stress test our system so that particularly complex queries, malformed queries, or malicious attacks are unlikely produce undesirable responses. A group of clinical AI specialists and domain experts comprise our "Red Team" who, with each major release, test how the system behaves in response to complex clinical scenarios, malformed queries, multi-turn conversations, and other attempts to provoke undesirable responses. Common targets for Red Team testing include biased responses, provocation of model knowledge, failure to correct errors in queries, and loss of context or failure to switch context in long conversations.

To date, the Red Team has conducted more than **1,000 attempts** to provoke undesirable responses, with more than **200 hours dedicated to adversarial testing**. The findings from this effort are used to improve system behavior, reduce the likelihood of harmful outputs, and codify known risks into rubrics for ongoing surveillance testing. This approach improves the predictability and quality of answers, hardens the UpToDate Expert AI system against malicious attack and unintended use, identifies underlying content issues that prompt edits, and improves the experience of use.

After testing methods to identify harm at scale, we concluded that assessment of harm is best measured by experts. Accordingly, we are developing an audit process that relies on our expert network of authors and section editors to identify potential harm in answers to user queries. With the deep domain knowledge of our experts, the goal of this process is to identify potentially meaningful inconsistencies that are difficult to capture with automated scoring alone.

Across these domains, automated measures are supplemented by structured human review, physician calibration, platform testing, and expert audit to connect scaled measurement back to clinical judgment. These measures are most powerful in combination. Each looks at a different risk in the AI-generated content clinicians read. Together, they allow us to evaluate the meaning of the words on the page — not just model performance on a benchmark, but the content a clinician may read, interpret, and use in patient care.

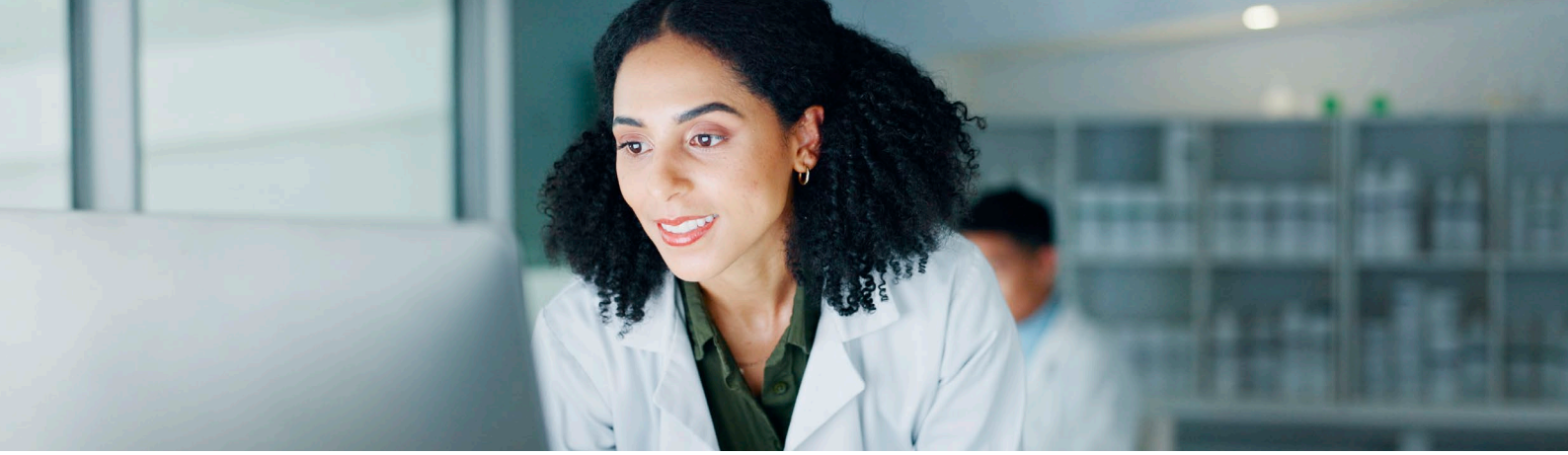
Scope and interpretation of comparator findings

We recognize that purpose-built clinical AI products and clinician-facing medical search tools are increasingly available. The comparator analysis reported here evaluated UpToDate Expert AI against two high-performing general-purpose LLMs using the same clinical queries and physician-authored rubric criteria. Purpose-built clinical AI products were not included in this analysis because their terms of use restrict this type of comparative evaluation. That distinction does not lessen the importance of these findings, but rather clarifies their scope.

The results of the comparator analysis demonstrate UpToDate Expert AI's performance under a validation method intended to assess point-of-care response quality, including what is present, what is omitted, and whether the response aligns with clinician-defined expectations. Any AI system that may influence patient care should be held to this level of rigor: evaluation using meaningful criteria, real-world use cases, and expert review to detect omissions, unsupported claims, misplaced confidence, loss of context, inappropriate assumptions, and other potential failure patterns that generic benchmarks may miss.



To date, the Red Team has conducted more than **1,000 attempts to provoke undesirable responses**, with more than **200 hours dedicated to adversarial testing**.



Expert judgment at scale

Expert judgment is built into the evaluation system for UpToDate Expert AI. The people defining, testing, and reviewing UpToDate Expert AI are not generic reviewers. They include UpToDate physician editors, clinical AI specialists, and selected members of the author and editor network who understand both the relevant clinical domain and the UpToDate content that grounds the answer.

That expertise matters. Evaluating AI-generated clinical content requires knowing what a clinician is trying to accomplish, which details matter in context, which omissions could change interpretation, and whether an answer is faithful to the underlying source content. Those judgments depend on clinical expertise and familiarity with the knowledge base.

Human experts in the loop help define what strong answers should include, build and maintain rubrics, calibrate automated grading, review real-world outputs, and identify where answers may be incomplete, unsupported, or potentially harmful. Clinical specialists and domain experts also lead red teaming for high-risk scenarios and known weaknesses in generative AI.

The system uses technology to scale evaluation, but expert judgment defines the standard and guides what should improve next.

Why other common AI evaluations are not enough

When generative AI entered clinical environments, the industry reached for familiar measurement tools: standardized medical licensing exams, published case challenges, user ratings, and general-purpose accuracy benchmarks. These tools can provide useful signals, but they were not intended to determine whether AI-generated content is appropriate for patient care (Table 1).

Test Type	Test Process	Pitfalls in Measuring Performance
Medical knowledge benchmarks	Whether a system can answer multiple-choice-like questions	• Non-native function of point-of-care CDS • Rarely updated along with medical evidence or practice • Falsely high passing rates due to LLMs training on the answers
User ratings	Whether an answer felt useful, clear, or efficient to the user	• Important to identify cases for analysis, but not a gold-standard for testing the reliability at scale
Clinical vignettes and case challenges	How a system performs on selected cases, scenarios, or teaching examples	• Adaptation of usual function to meet test-specific tasks • Not updated along with medical evidence or practice

Table 1: *Common AI validation methods*

The UpToDate advantage

For more than 30 years, UpToDate has helped clinicians answer questions at the point of care through expert-authored, peer-reviewed, continuously updated clinical content. With UpToDate Expert AI, that responsibility extends to AI-generated answers: helping clinicians access information that is current, reliable, and useful at the point of care.

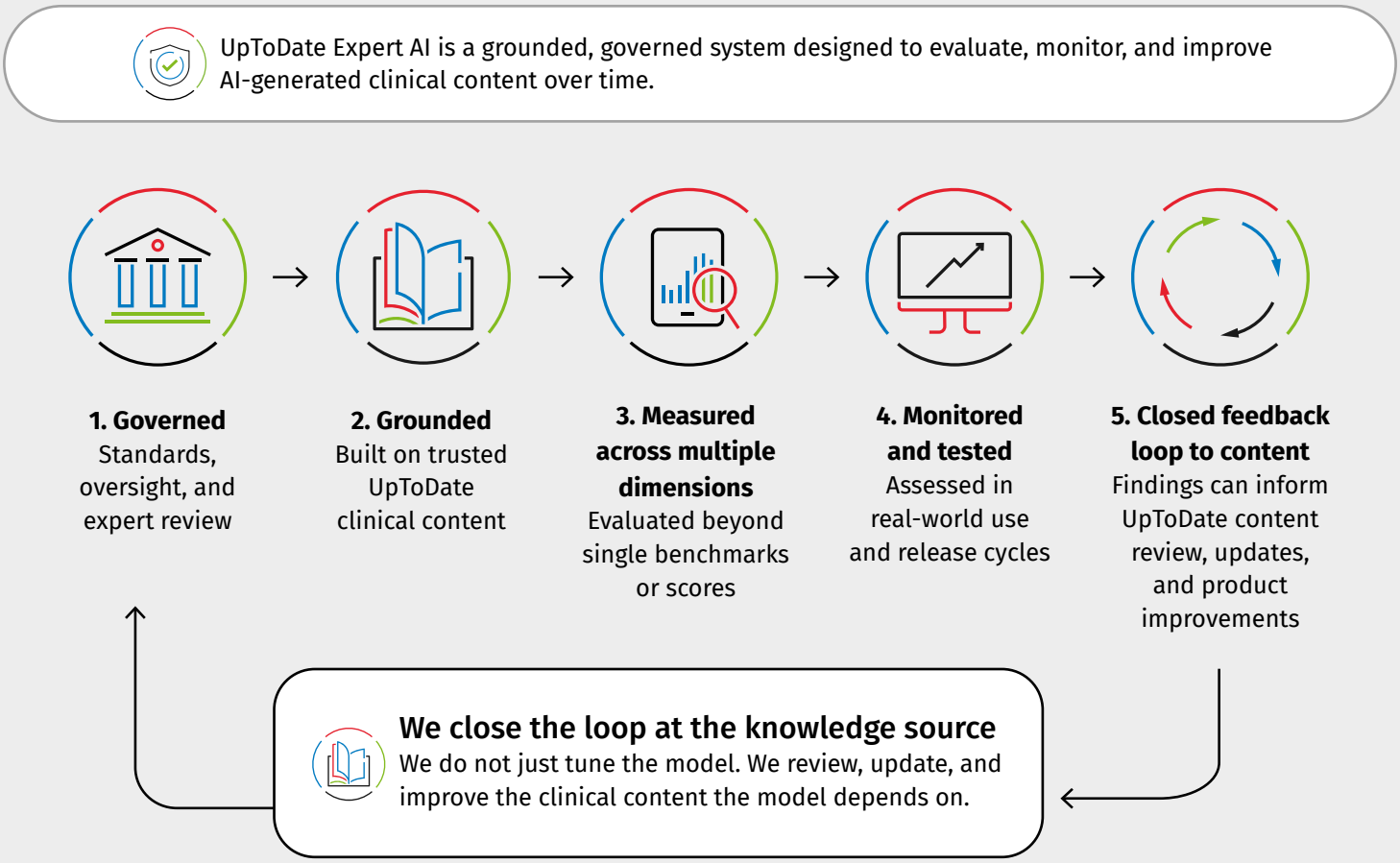


Figure 4: *UpToDate Clinical Intelligence Model*

UpToDate Expert AI extends that foundation into AI-generated content. It is not just a model producing answers to help guide care decisions. It is part of a governed system built around trusted clinical knowledge, clinician-defined standards, expert review, knowledge-integrity checks, red teaming, product guardrails, monitoring, and feedback loops to the underlying content and our AI system (Figure 4).

This is where UpToDate's experience matters. The same principles that have guided UpToDate for decades — expert authorship, editorial rigor, clinical relevance, evidence, and continuous updating — now shape how UpToDate Expert AI is evaluated and improved.

For AI-generated clinical content, reliability depends on the system around the model: how it is grounded, governed, evaluated, monitored, and continuously improved.

©2026 UpToDate, Inc. and its affiliates and/or licensors. All rights reserved.