Counsel

LLM-as-a-Judge:

How AI can assess clinical quality in asynchronous care at scale

LLM-as-a-Judge: How AI can assess clinical quality in asynchronous care at scale

As care delivery shifts toward AI-enabled models, patient safety has become a critical evaluation criteria for healthcare organizations. One question that increasingly defines enterprise adoption is, how do you measure clinical quality at scale? With one in three Americans turning to unsupervised consumer AI tools for health information[1], organizations face growing urgency to deploy medical AI to expand access. To do so responsibly, they must ensure these models have rigorous clinical quality measures in place to safely deliver care at scale.

For decades, Clinical Quality Assurance (CQA) has relied on manual peer review, having physicians auditing charts against established rubrics composed of evidence-based clinical protocols. That model, while rigorous, was built for a world where care volumes were finite and interactions were episodic. Today, AI-enabled care platforms generate thousands of interactions, or threads, per week across both AI-only and physician-led chats. Traditional random sampling methods struggle to keep pace with this growth.

For healthcare organizations across the continuum, this creates a new category of risk. AI can expand access and reduce clinical strain, but only if its outputs provide the appropriate degree of oversight.

To address this challenge, Counsel Health's AI Research and Clinical teams partnered to design and implement an automated LLM-as-a-Judge framework. Rather than relying solely on humans to retrospectively review a small subset of interactions, this approach leverages specialized AI judges that assess each thread against standardized, clinician-defined rubrics. This framework enables organizations like Counsel to improve alignment around evidence-based protocols, detect safety deviations with higher sensitivity, reduce reviewer fatigue and operational bottlenecks, and quantitatively evaluate AI and physician performance across defined metrics.

In this report, we'll examine

- How an LLM-as-a-Judge framework can be implemented across clinical domains
- The efficacy of LLM judges against human physician review
- What this new framework means for defensible governance when medical AI is deployed at scale

[1] KFF. (n.d.). Poll: 1 in 3 adults are turning to AI chatbots for health information, equaling the share who use social media for health. https://www.kff.org/health-information-trust/poll-1-in-3-adults-are-turning-to-ai-chatbots-for-health-information-equaling-the-share-who-use-social-media-for-health/

Factors driving this analysis

CQA is the bedrock of safe healthcare delivery and process improvement, yet it remains one of the most challenging operational bottlenecks in modern medicine. Traditional CQA relies on clinical peer review, a process where clinicians manually audit a subset of patient charts to ensure guideline adherence. As healthcare shifts increasingly toward asynchronous and AI-enabled care models, the volume of clinical data to review has exploded. This growth has led to significant challenges in traditional manual CQA methods for two reasons:

Operational fatigue

Clinical leaders are forced to choose between devoting providers' time toward conducting CQA versus expanding their coverage of services. CQA requires a significant time investment as physicians have to be trained to follow nuanced clinical logic present in rubrics.

Reviewer fatigue

Clinicians tasked with reviewing large batches of threads are subject to reviewer fatigue, which could lead them to miss "black swan" low-frequency, high-severity safety deviations.

Counsel experienced a similar explosion in volume of AI-only and physician-led threads. As a result, relying purely on traditional CQA was no longer a viable tool within its risk management strategy. To ensure patient safety at scale, Counsel designed an automated AI-powered system that addresses both of the aforementioned challenges.

The organization's motivation for developing an automated LLM-as-a-Judge framework is rooted in a simple premise: heightened clinical auditability leads to increased patient safety. In a world where medical AI leads to unlimited clinical capacity, every thread should be evaluated for its clinical rigor and guideline adherence by an impartial physician judge who follows a standardized rubric.

Counsel sought to replicate this oversight by establishing a repository of standardized LLM judges, each independently applied to specific threads to detect potential deviations from evidence-based protocols. For example, a binary LLM judge can assess whether a clinician documented pregnancy status for female patients under the age of 50 presenting with a urinary tract infection. This check is critical to prevent the prescription of contraindicated medications and ensure adherence to guideline-based care.

To design these LLM judges, Counsel aimed to standardize and formalize implicit criteria that physicians were already applying during CQA. This required decomposing existing CQA workflows into multiple targeted LLM judges, such as a UTI judge that assessed pregnancy status, screened for pyelonephritis symptoms, and checked for contraindicated medications, followed by aggregating the outputs into a composite score.

Implementation method and data collection

We adopted an LLM-as-a-Judge framework to automate two CQA workflows: antibiotic stewardship for viral upper respiratory infections and acute sinusitis, and condition-specific audits designed around urinary tract infections (UTIs) and vaginitis.

These two workflows were chosen because they impact the largest volume of physician-led threads at Counsel. Antibiotic stewardship is important due to the high volume of patients that present with viral flu symptoms, especially during the late-winter flu season. Condition-based audits are important to ensure adherence to evidence-based guidelines when prescribing medications for UTIs and vaginitis.

Antibiotic stewardship

We started with a high-volume dataset containing 6,000+ physician-led threads and 136,000+ discrete messages from January to February 2026, which covers a representative cross-section of patient interactions across Counsel's asynchronous care delivery platform. For antibiotic stewardship, the research team focused specifically on viral upper respiratory infections (URI) and acute sinusitis, as these represent high-volume conditions where antibiotic over-prescription remains a significant public health challenge, and where manual audits are most likely to fail due to their sheer volume.

Designing the automated viral URI LLM judge

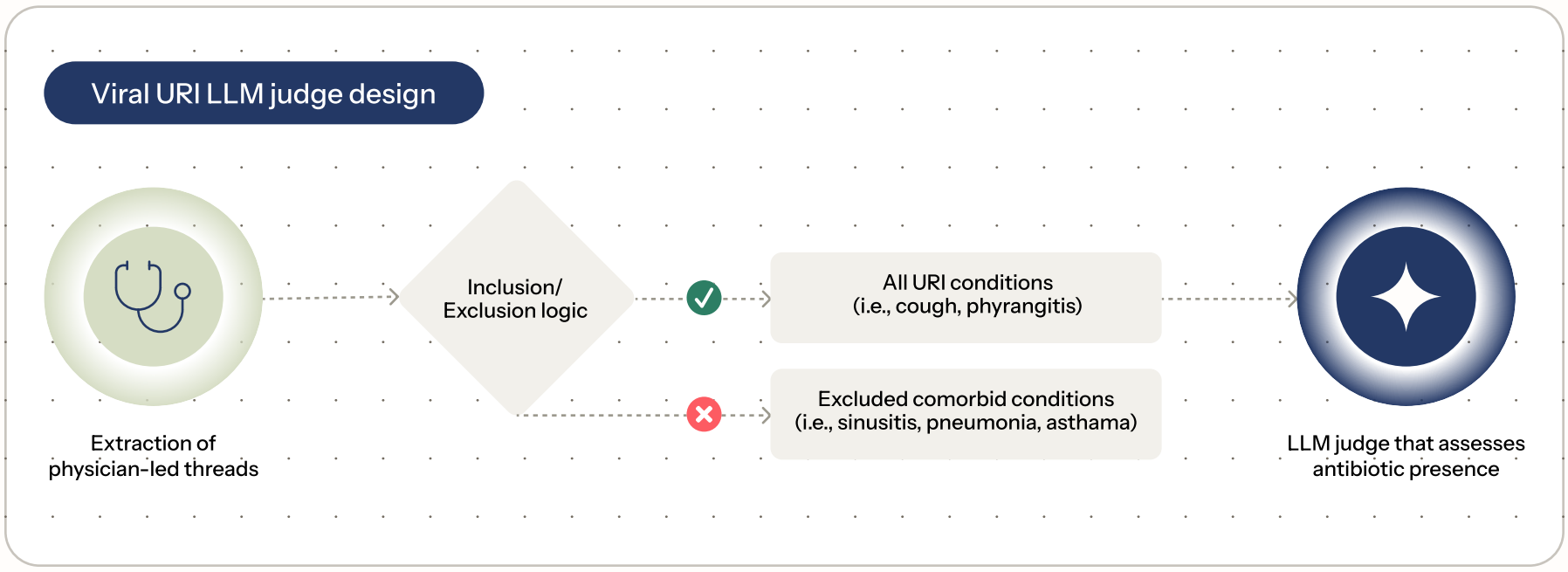
The viral URI LLM judge reviews threads to identify when a physician inappropriately prescribes antibiotics for a viral URI, which are a set of conditions that are usually not treated with antibiotics. The judge reads messages from the thread, the physician-generated note, and the set of associated prescription orders to assess if an antibiotic was prescribed. The workflow additionally uses regex filtering across physician-assigned diagnoses to programmatically exclude clinical confounders, such as pneumonia and asthma.

6K+

Starting dataset of physician-led threads between January and February 2026

136K+

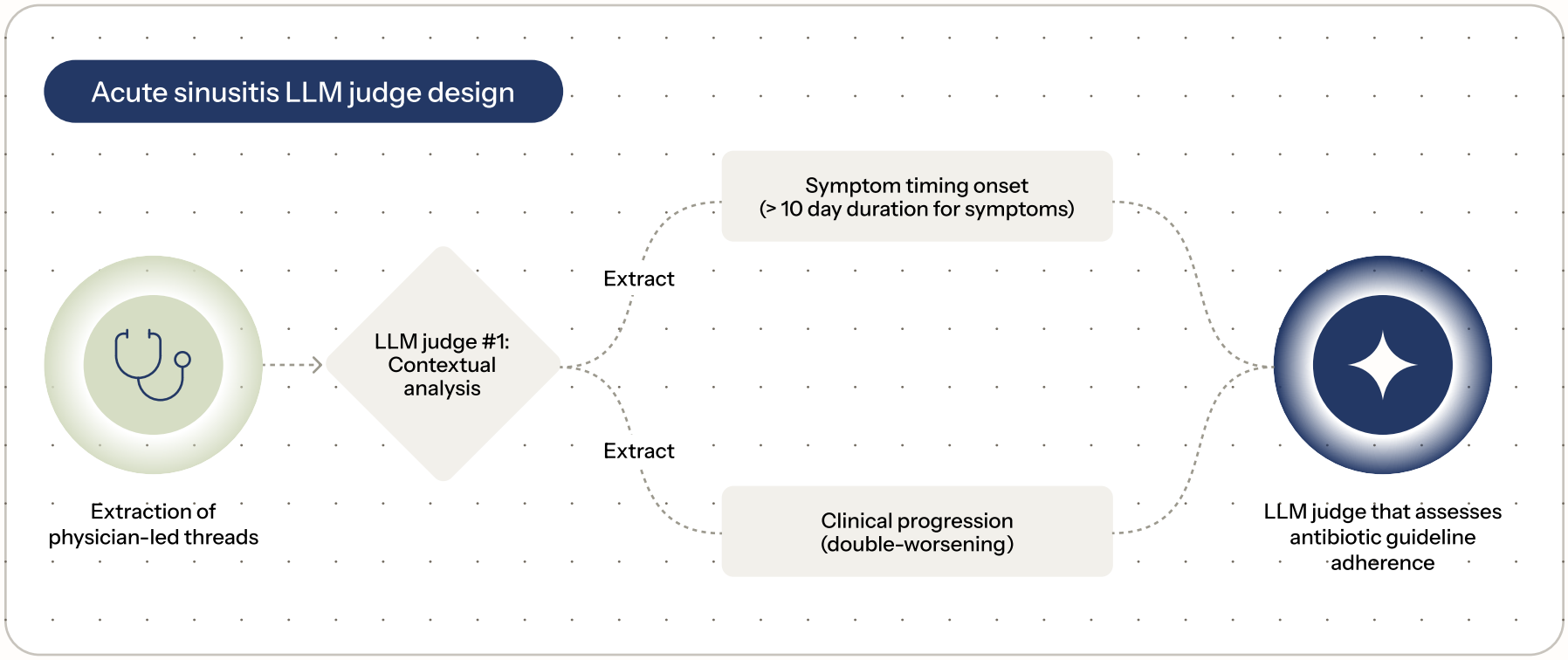
Discrete messages within the dataset of threads between January and February 2026



Designing the automated acute sinusitis LLM judge

The acute sinusitis LLM judge assesses the appropriateness of antibiotics prescribed for sinusitis, a condition where antibiotics are often contraindicated. Unlike the viral URI LLM judge, the acute sinusitis judge was also asked to output particular criteria to contextualize its decisions, such as calculating the symptom onset timing (i.e., if symptoms persist for more than 10 days, antibiotics could be appropriate), as well as clinical trajectory (i.e., whether there was improvement followed by a sudden worsening of symptoms, also known as “double-worsening”, which could indicate that antibiotics are appropriate).

After contextualization with the symptom onset timing and the double-worsening criteria, the judge is asked to determine if appropriate clinical guidelines were followed.



Condition-specific audits

Building on the antibiotic stewardship judges, we expanded our framework to conduct condition-specific audits for conditions frequently treated on Counsel’s platform. We particularly focused on UTI and vaginitis, due to known contraindications associated with standard treatments and commonly prescribed medications.

Automated UTI and vaginitis LLM judge

For these higher-complexity workflows, we developed a suite of seven independent binary LLM judges.

Each of these binary judges reviews threads to assess a specific clinical safety criteria, such as whether the clinician inquired about pregnancy and breastfeeding status, or previous UTI history. A composite score was then assigned by aggregating the individual independent LLM outputs based on a rubric generated by the clinical team. There were four possible composite scores or labels: excellent, acceptable, inadequate, or a near miss.

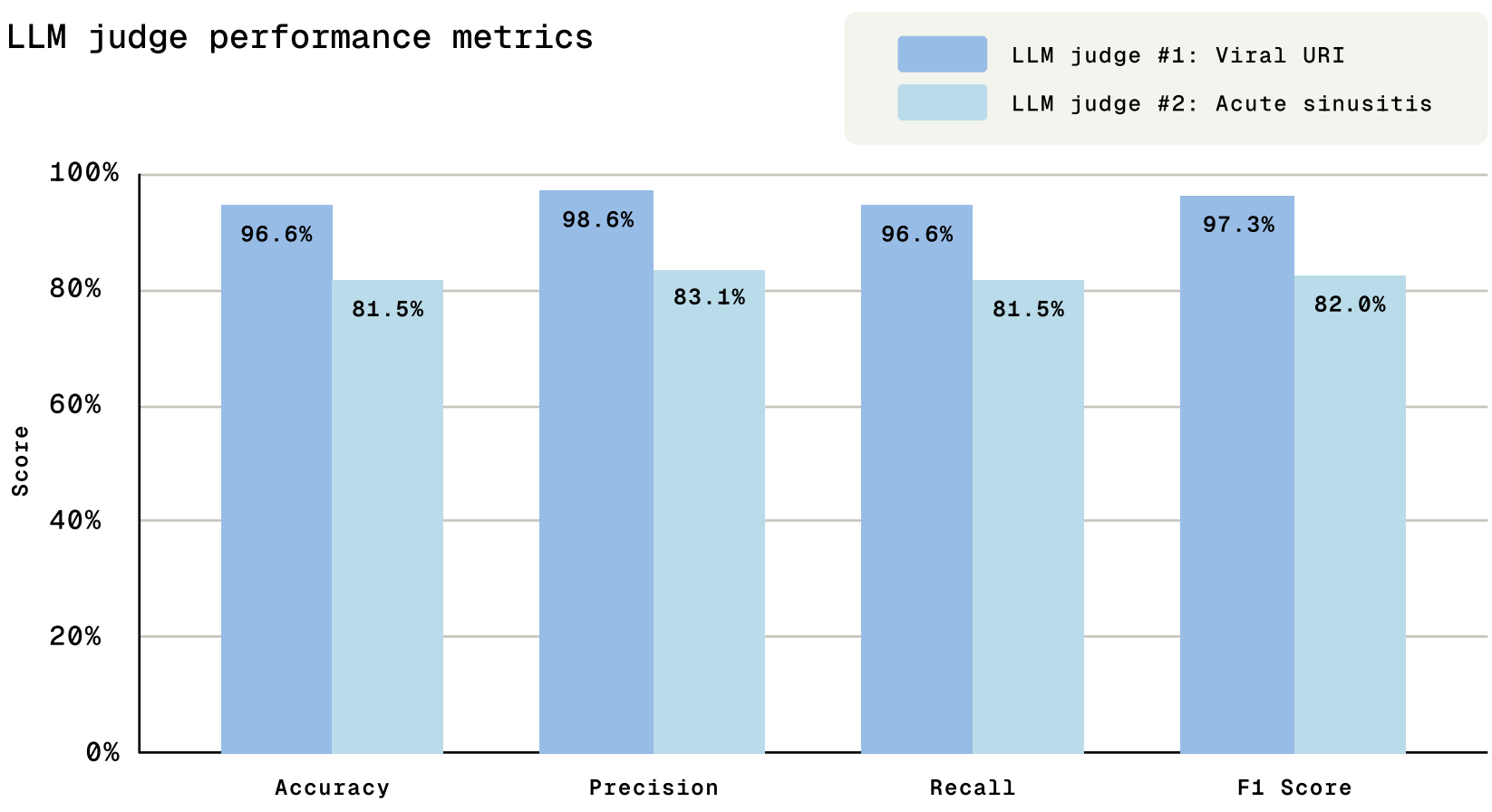
Results from the data analysis

Antibiotic stewardship results

We evaluated the performance of the LLM judges against a subset of “gold standard” threads labeled by human physicians from January to February 2026. The viral URI judge was 96.59% accurate in detecting inappropriate antibiotic prescriptions compared to a human judge. It had a weighted precision of 98.6% and a weighted recall of 96.6%, meaning that the judge rarely flagged appropriate care as an error (high precision), while remaining sensitive to the majority of incorrect prescribing instances (high recall). The F1 score, which considers both the precision and recall together as a metric, was 97.3%, while Cohen's Kappa of 0.557 indicates moderate to strong levels of agreement between the LLM judge and physician.

The acute sinusitis judge achieved an 81.52% accuracy and an F1 score of 82%, reflecting a balanced performance between its 83.1% weighted precision and 81.5% weighted recall. A Cohen’s Kappa of 0.551 indicates a moderate level of agreement, suggesting that while the judge effectively identifies inappropriate prescribing patterns, the clinical complexity of sinusitis results in more frequent disagreement with human physicians than seen in the URI subset.

	LLM judge #1:Viral URI	LLM judge #2: Acute sinusitis
Total threads evaluated (N)	230	92
Accuracy (%)	96.59%	81.52%
Precision (Weighted)	98.6%	83.1%
Recall (Weighted)	96.6%	81.5%
F1 Score (Weighted)	97.3%	82.0%
Cohen’s Kappa (κ)	0.557	0.551



The accuracy, precision, recall, and F1 statistics indicate that the viral URI LLM judge aligns with the physician judges across the same threads. The main disagreements occurred when the LLM judge identified additional mentions of antibiotics in the provider notes than the physician reviewers.

The acute sinusitis LLM judge remains a capable evaluator but encountered greater difficulty in distinguishing borderline cases compared to the URI judge. This performance suggests that the clinical nuances of sinusitis, such as duration of symptoms and “double worsening”, presents a more complex classification task for the LLM judge than the more straightforward viral URI criteria. When reviewing disagreements between the LLM judge and the physician labels, disagreements largely stemmed from differences in perceived symptom onset time. The LLM judge tended to calculate symptom duration based on the time from first complaint, whereas the physicians applied clinical intuition and excluded potentially irrelevant symptoms that may have been mentioned across the chat.

Condition-specific audit results

We similarly evaluated the performance of the condition-specific LLM judge for UTI and vaginitis against a “gold standard” rubric created by human physicians from January to February 2026. Overall, the LLM judges were harsher graders than physicians. Threads marked as “excellent” or “acceptable” by the LLM judges were similarly marked by physicians. For threads that performed poorly across various rubric criteria, physicians were more likely to be lenient. For example, when we looked at threads marked as “inadequate” or “near miss” by the LLM judges, a physician noted that the thread “technically did not ask about pregnancy in a 47 year old, but given the upgrade [in intensity, the physician] did not dock for that”.

Table 1 - cross-tab results for the UTI and vaginitis condition-specific audits. Threads marked as “excellent” or “acceptable” by the LLM judges were similarly marked by physicians. Physicians were more likely to be lenient, and assigned fewer “inadequate” or “ near miss” labels.

What LLM judges unlock for the future of patient care

Impact to prospective patient care

Beyond retrospective audits, these findings pave the way for deeper, more integrated safety checks within Counsel's Clinician Cockpit, our proprietary care delivery interface, reinforcing our commitment to high-quality, evidence-based care.

We are actively embedding these LLM-as-a-Judge workflows into future iterations of our Clinician Cockpit, enabling real-time alerts before prescriptions are finalized. By shifting from retrospective quality review to prospective intervention, we can reduce inappropriate antibiotic utilization at the point of care, enhancing patient safety while optimizing clinician workflows in AI-enabled medicine.

Impact on clinical capacity

The implementation of an LLM-as-a-Judge framework expands clinical capacity at a time when the U.S. is facing significant physician shortages. By saving providers thousands of hours per year in manual CQA, clinicians can now focus on care delivery, particularly impactful for complex conditions that can be treated upstream before escalating to high-cost settings.

It takes a Counsel physician approximately an hour to review 20 threads for a single workflow. For conditions such as viral URI, a manual CQA process would require physicians to spend over 200 hours each month or 2,400+ hours per year. When multiplying these review hours across multiple conditions, it is clear that manual CQA becomes prohibitive.

As we expand our LLM judges across more conditions, we will automate repetitive audit tasks and empower Counsel physicians to focus on care delivery rather than manual rubric adherence reviews. These LLM judge workflows create a repeatable, protocol-driven system for CQA at scale, one designed to uphold strict standards while potentially saving human physicians millions of hours of review time over the next five years as thread volumes and clinical complexity continue to grow.

Expanding care access, responsibly.

For enterprise stakeholders, the implications of this report are broader. AI-enabled care models promise expanded access, reduced total cost of care, decreased strain on clinical capacity, and improved network efficiency. To safely deploy medical AI at scale, however, defensible governance will be critical.

The ability to evaluate every interaction, quantify performance using precision, recall, and F1 scores, detect safety deviations with high sensitivity, reduce variability introduced by reviewer fatigue, and align clinical outputs to standardized, evidence-based protocols signals operational maturity to responsibly serve large, diverse patient populations.

For health plans, these capabilities introduce stronger risk management and improved alignment between care delivery and governance standards. For health systems, it enables scalable clinical quality oversight without increasing administrative burden. For consumer health platforms, it establishes a measurable safety infrastructure that builds user trust.

This report demonstrates that medical AI can do more than deliver care, it can be leveraged to design evaluation systems that strengthen accountability across every patient interaction. As AI-enabled care models become integrated into the front door of care, LLM-as-a-Judge frameworks will provide the clinical visibility required to responsibly scale access.

Request a demo

See how Counsel's responsible AI-enabled care model delivers measurable impact across your clinical strategy.

counselhealth.com/contact-sales | business@counselhealth.com

Try Counsel now
Scan the QR code to download the app and start your first chat.