SIGNAL 1

INSIGHTS PAPER

LLM-as-a-Judge in Healthcare

As healthcare organizations deploy generative AI across clinical, operational, and administrative workflows, a new challenge is emerging: **how to evaluate AI-generated output at scale.**

A patient messaging application may generate hundreds of thousands of responses each year. Ambient documentation tools can create clinical notes for every patient encounter. Clinical copilots may provide recommendations dozens of times per provider, per day. The volume of AI-generated content quickly exceeds what can be reviewed manually.

This challenge is driving growing adoption of **LLM-as-a-Judge**: using a language model not to generate content, but to evaluate it against defined criteria such as policy adherence, factual accuracy, completeness, clinical appropriateness, and tone.

Why it works

Generating content and evaluating content are fundamentally different problems.

When a generative model **produces output**, it is solving an open-ended task with an effectively unlimited number of possible responses.

When an evaluator reviews that output, it is performing a bounded reasoning task against explicit criteria:

- Does the output follow organizational policies?
- Does it contain unsupported claims?
- Is important information missing?
- Does it meet expected quality standards?

This distinction is sometimes described as **verification asymmetry**: it is fundamentally easier to verify information than to create it.

Put simply, writing the essay is harder than grading it against a rubric.

Who monitors the judge?

Evaluator models are not exempt from governance. Like any other production AI system, they must be monitored and validated over time.

The most important mechanism is **human calibration**. Much like quality assurance programs periodically audit clinical documentation or coding decisions, organizations should periodically compare the LLM judge's evaluations against expert human review of the same outputs. If the evaluator consistently identifies the same issues and reaches similar conclusions, confidence increases that it is applying organizational standards appropriately. When meaningful disagreement emerges, the evaluation criteria, prompts, or underlying model can be refined.

The goal is not to create an evaluator that is blindly trusted. The goal is to create an evaluation process that is continuously measured against human judgment and improved over time.

The emerging verification layer

As generative AI adoption accelerates, healthcare organizations need a scalable way to monitor quality and performance in production.

LLM-as-a-Judge is emerging as a critical verification layer — not because it replaces human oversight, but because it makes human oversight scalable.

Ultimately, the answer to "Who watches the AI?" is not another model.

It is an AI management system that ensures every component — including the judge itself — remains observable, measurable, and accountable.