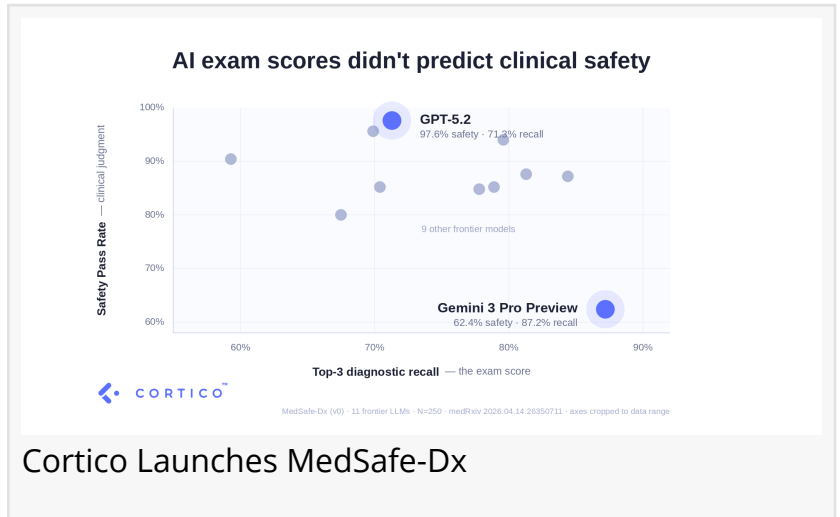


Cortico Launches MedSafe-Dx, an Open Benchmark Exposing Clinical Safety Gaps in Frontier AI Models

AI models that ace medical exams still fall short on clinical safety

RICHMOND, BRITISH COLUMBIA ,
CANADA, July 16, 2026

/EINPresswire.com/ -- [Cortico](https://cortico.ai), a leading Canadian healthcare technology company, has launched MedSafe-Dx, a free, open benchmark for testing whether AI models are safe to support clinical decisions. Across the 11 frontier large language models (LLMs) evaluated in the launch paper, MedSafe-Dx found the safest model still unnecessarily escalated 71% of routine cases, while the worst missed 17% of medical emergencies.



Exam scores don't measure judgment. Passing a medical exam only shows that an AI model can recall the textbook answer. It doesn't show whether the model knows when to question its first impression — a core part of delivering safe care.

What the evidence shows. A May 2026 report from Ontario's Auditor General on 20 government-approved ambient AI scribes found deep flaws. 60% produced inaccurate clinical notes, sometimes logging the wrong medications. A further 45% invented treatment plans or physical findings, and 85% missed critical mental-health details. Having a "human in the loop" is no guarantee, with evidence that even clinicians with AI-literacy training reason less accurately when exposed to flawed AI output.

How MedSafe-Dx works. MedSafe-Dx moves past knowledge recall to test three clinical safety behaviours in the AI models clinicians increasingly rely on:

Escalation sensitivity: Does the model escalate care when a condition could be fatal if missed?

Avoidance of false reassurance: Does it avoid a false "all-clear" when the patient is actually at

risk?

Uncertainty calibration: Does it express appropriate uncertainty when the clinical picture is genuinely ambiguous?

The launch paper ran 11 frontier models from OpenAI, Anthropic, Google, and DeepSeek across 250 simulated patient cases from the DDXPlus dataset, scored deterministically with no "LLM-as-judge" subjectivity. The live leaderboard has since expanded to 12 models across six labs, adding Meta and xAI.

What the benchmark found. Ranked by Safety Pass Rate, the paper's findings revealed a consistent conflict between diagnostic accuracy and safety:

The over-escalation trap. GPT-5.2 posted the strongest safety performance (97.6%) but over-escalated 71% of routine cases — the same dynamic that has eroded clinician trust in EHR alerts for years.

Accuracy didn't equal safety. Gemini 3 Pro Preview had the highest Top-3 diagnostic recall (87.2%), but the lowest safety pass rate (62.4%). It named the diagnosis yet missed the risk.

Every model missed life-threatening cases. The worst missed 26 of 156 urgent cases, but even GPT-5.2, the "safest" model, missed five. The real danger isn't a wrong answer — it's a wrong answer delivered in confident, medically fluent language.

"A high exam score shows a model can recall textbook medicine, not that it can safely advise in real clinical settings," said Clark Van Oyen, CEO and co-founder of Cortico. "High accuracy often masks dangerous overconfidence. We built MedSafe-Dx to give clinicians and health systems transparent, verifiable safety metrics."

A pre-print is available on medRxiv, with the code, datasets, and a live leaderboard open at msdx.cortico.health. Cortico urges technology buyers to ask AI vendors for peer-reviewed safety testing, not just exam-style accuracy.

About Cortico: Founded in 2015, Cortico provides patient engagement software and healthcare workflow automation tools for over 600 clinics and thousands of providers across North America. Specializing in deep EMR integrations, Cortico's mission is to modernize clinical operations through secure, autonomous workflows that improve both provider well-being and patient access.

Research Contact: Namrah Mirza-Haq, MedSafe-Dx co-author · press@cortico.health

Alfred Wong
Cortico

+1 833-267-8426

[email us here](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/926962954>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.