MedSafe-Dx (v0): A Safety-Focused Benchmark for Evaluating LLMs in Clinical Diagnostic Decision Support

Clark Van Oyen[1], Namrah Mirza-Haq[2]

[1] *Cortico Health Technologies*

[2] *Ziya Analytics*

Abstract

MedSafe-Dx (v0), introduces a new safety-focused benchmark for evaluating large language models in clinical diagnostic decision support using a filtered subset of the DDx Plus dataset (N=250). MedSafe-Dx evaluates three dimensions: escalation sensitivity, avoidance of false reassurance, and calibration of uncertainty. Models were tasked with providing a ranked differential (ICD-10), an escalation decision (Urgent vs. Routine), and a confidence flag. Performance was measured via a "Safety Pass Rate," a composite metric penalizing three hard failure modes: missed escalations of life-threatening conditions, overconfident incorrect diagnoses, and unsafe reassurance in ambiguous cases. Eleven models were evaluated and revealed a significant disconnect between diagnostic recall and safety. GPT-5.2 achieved the highest Safety Pass Rate (97.6%), while several models exhibited high rates of missed escalations or unsafe reassurance. MedSafe-Dx provides a robust stress test for identifying high-risk failure modes in diagnostic decision support and shows that high diagnostic accuracy does not guarantee clinical safety. While the benchmark is currently limited by synthetic data and proxy labels, it provides a reproducible, auditable framework for testing AI behavior before clinical deployment. Our findings suggest that interventions such as safety-focused prompting and reasoning-token budgets could be essential components for the safe deployment of LLMs in clinical workflows.

1. Introduction

The integration of Large Language Models (LLMs) into clinical workflows has transitioned from theoretical exploration to active implementation in clinical decision support (CDS) systems [1,2]. Physicians face ever-increasing workloads as healthcare worker shortages persist [3]. In addition to many hospitals being understaffed, physicians report high levels of burnout due to administrative burden [4]. There is great need for tools to streamline processes and magnify the capabilities of physicians to address unmet need in the healthcare space. The promise of LLMs to enhance healthcare delivery is contingent upon their ability to provide reliable diagnostic assistance and triage automation, effectively serving as a 'force multiplier' that mitigates the risks associated with physician fatigue and systemic understaffing.

However, the rapid adoption of these technologies has outpaced the development of clinically meaningful safety benchmarks. Some existing benchmarks evaluate LLM safety based upon diagnostic recall and medical knowledge accuracy such as MedQA. Others have a clinical reasoning focus such as MedAlign or HealthBench. Clinical reasoning benchmarks are an integral part of evaluating LLM safety in healthcare settings as they simulate real conversations between patients and providers and help evaluate the use-cases of LLMs in real-world settings. Our benchmark goes further along this vein and measures the 'judgement' capabilities of LLMs. We evaluate whether a model can correctly escalate a case. It is crucial for LLMs not only to be

diagnostically accurate, but to also correctly identify and escalate crisis if they are to be integrated into healthcare workflows.

For instance, studies have shown that while frontier models can achieve over 90% accuracy on static medical questions, they often fail to identify critical "red flag" symptoms in simulated triage, leading to potentially catastrophic delays in care [5,6]. Furthermore, research into LLM calibration indicates that models frequently express high verbal confidence even when their underlying diagnostic logic is flawed, a phenomenon known as "unsafe reassurance" [7]. These findings suggest that current benchmarks, which aggregate performance into a single accuracy metric, may inadvertently mask high-risk failure modes that render a model unsuitable for real-world clinical deployment [8].

In this study, we introduce MedSafe-Dx, a deterministic safety stress test designed to evaluate these specific failure modes. By utilizing a rules-based evaluation of frontier LLMs, we demonstrate that diagnostic proficiency is not always correlated with clinical safety [8]. Our findings suggest that without specific architectural safeguards, even high-performing models can provide "unsafe reassurance" or miss critical escalations, potentially exacerbating the very risks that AI-augmented workflows aim to reduce. We designed MedSafe-Dx to act as a safety stress test for diagnostic decision support. Rather than measuring knowledge breadth, we ask three specific safety questions:

- **Escalation sensitivity:** Does the model escalate care when a condition could be fatal if missed?
- **False reassurance:** Does it avoid telling a patient they are fine when they are actually at risk?
- **Calibration:** Does it express appropriate uncertainty when the clinical picture is genuinely ambiguous?

We operate from the principle that for clinical decision support tools, safety is a prerequisite for utility, not a variable to be traded off against accuracy. A model with perfect diagnostic recall but frequent missed escalations is dangerous.

2. Materials

The MedSafe-Dx (v0) benchmark was constructed using a reproducibly sampled subset (N=250, Seed 42) of the January 2026 DDx Plus dataset [9]. DDx Plus is a large-scale synthetic clinical dataset comprising 1.3 million patient profiles generated from a proprietary medical knowledge base and a rules-based diagnostic engine. Each profile includes a ground-truth pathology, a complete differential diagnosis with associated probabilities, and a comprehensive set of symptoms and medical antecedents.

The architecture of DDx Plus is integral to this benchmark because every pathology is assigned a severity score (1–5), where lower values denote higher clinical acuity. This structured metadata provides an objective, deterministic environment to assess triage safety, mimicking the nuanced data a physician must synthesize during a patient encounter.

3. Methods

3.1. Gold Label Extraction and Triage Logic

For each sampled case, we extracted three components to form a "Gold Label" standard:

- **Differential Mapping:** The top-3 differential diagnosis codes from DDx Plus were extracted and mapped to their respective severity scores.
- **Deterministic Triage:** A binary triage label was assigned based on acuity. If any of the top-3 diagnoses featured a severity score $1 \leq 2$, the case was labeled ESCALATE_NOW; otherwise, it was labeled ROUTINE_CARE.
- **Ambiguity Flagging:** Cases were flagged as "Ambiguous" if the ground-truth probabilities between the primary diagnosis and high-risk alternatives were within a narrow margin, or if the case presented conflicting severity markers.

3.2. Model Inference

Frontier LLMs were presented with the structured patient symptoms and antecedents (excluding gold labels). Models were instructed to provide a response in a standardized JSON format including: (1) a differential of 5 ICD-10 codes, (2) a triage decision (ESCALATE_NOW or ROUTINE_CARE), and (3) a confidence self-assessment (CONFIDENT or UNCERTAIN).

3.3. Deterministic Evaluation and Safety Pass Rate

Rather than utilizing "LLM-as-a-judge" or subjective rubrics, we utilized a Python-based scoring engine to compare model outputs against the DDx Plus gold labels. Models that failed to adhere to the output schema were marked as Coverage Failures (0% score). The engine assessed cases for three "Hard Fail" safety violations:

- **Missed Escalation:** The gold label required urgent care, but the LLM selected routine care.
- **Unsafe Reassurance:** The case was marked as ambiguous or high-risk, but the LLM selected ROUTINE_CARE with CONFIDENT status.
- **Overconfident Wrong:** The LLM expressed CONFIDENT status but failed to include the ground-truth diagnosis in its top-3 ICD-10 codes.

Accuracy was treated as a secondary metric: Top-3 Diagnostic Recall was only calculated for cases that first passed all safety checks. The primary metric, the Safety Pass Rate, represents the percentage of cases that resulted in zero hard safety failures, ensuring that the benchmark prioritizes patient safety over simple knowledge retrieval. Models may achieve a perfect safety score by simply escalating every case. This is by design, as it allows the healthcare professional to have visibility and control of the trade-offs between over-escalation and missed safety issues, and choose models appropriate for their use case. It is worth noting, given the large-scale availability of the DDx Plus dataset, it is possible that evaluated models encountered portions of this data during pre-training. However, the MedSafe-Dx benchmark remains valid because we are measuring behavioral safety rather than simple knowledge recall. The benchmark evaluates the model's ability to avoid false reassurance. Even if a diagnosis is correct, the model must express appropriate uncertainty in ambiguous cases to pass the safety check. In addition, appropriate escalation requires the model to recognize when an identified diagnosis constitutes a clinical emergency requiring immediate intervention versus routine follow-up, requiring the model to show reasoning capabilities beyond diagnostic recall.

3.4. Clinical Deployment Scenarios

To evaluate the model's versatility across the care continuum, we utilized two distinct prompting scenarios:

- **Urgent Care Priority Triage:** In the acute care setting, the model functions as a primary triage interface designed to maximize sensitivity for life-threatening pathologies and minimize the incidence of high-acuity under-triage.
- **Primary Care Chart Review:** The model acts as a safety net during retrospective chart review, flagging potential escalations or omissions that a clinician might have missed during a busy outpatient session.

3.5. Variability Analysis

To account for the non-deterministic nature of LLMs, we conducted a variability analysis by running a subset of 50 cases across three separate iterations for each model. This allowed us to measure "Decision Drift"—instances where the same patient history resulted in conflicting triage labels—calculating a Consistency Score that reflects the model's reliability for production environments.

3.6. Ethics Approval and Data Availability

This study used a filtered synthetic subset of the publicly available DDx Plus dataset. No real patient data or protected health information were used. Therefore, institutional review board approval was not required. Evaluation code and benchmark specifications are available from the corresponding author upon reasonable request.

4. Results

The evaluation of eleven frontier Large Language Models (LLMs) on the MedSafe-Dx benchmark reveals a notable disparity between medical knowledge and clinical safety. The results are summarized in Table 1, categorized by their primary safety metrics and secondary diagnostic accuracy.

Table 1. *Safety Pass Rate, failure mode counts, and diagnostic recall for 11 frontier LLMs on MedSafe-Dx (N=250). Models ranked by Safety Pass Rate. Note: Gemini 2.5 Pro and Gemini 2.5 Flash Lite excluded due to severe API issues (0–8% valid responses).*

Rank	Model	Coverage	Safety Pass Rate (95% CI)	Missed Esc.	Overconf. Wrong	Unsafe Reassur.	Top-1 Recall	Top-3 Recall	Escalated (urgent)	Over-escal. (non-urgent)
#1	GPT-5.2 (OpenAI)	100%	97.6% (94.8–99.0)	5	1	0	55.7%	71.3%	N/A	67
#2	Claude Haiku 4.5 (Anthropic)	100%	95.6% (92.3–97.6)	1	1	0	48.5%	69.9%	N/A	62
#3	GPT-5 Chat (OpenAI)	100%	94.0% (90.3–96.4)	8	6	1	60.0%	79.6%	N/A	54
#4	GPT-4o Mini (OpenAI)	93%	90.4% (86.0–93.6)	3	13	3	38.9%	59.3%	N/A	69
#5	GPT-4.1 (OpenAI)	100%	87.6% (82.8–91.2)	13	12	5	58.4%	81.3%	N/A	50
#6	Claude Sonnet 4.5 (Anthropic)	100%	87.2% (82.4–90.9)	18	7	8	63.3%	84.4%	N/A	56

Rank	Model	Coverage	Safety Pass Rate (95% CI)	Missed Esc.	Overconf. Wrong	Unsafe Reassur.	Top-1 Recall	Top-3 Recall	Escalated (urgent)	Over-escal. (non-urgent)
#7	GPT-OSS 120B (OpenAI)	100%	85.2% (80.2–89.2)	17	16	4	55.4%	78.9%	N/A	46
#8	DeepSeek Chat v3 (DeepSeek)	100%	85.2% (79.6–88.9)	18	10	10	51.6%	70.4%	N/A	57
#9	GPT-5 Mini (OpenAI)	88%	84.8% (79.6–88.9)	9	0	0	65.6%	77.8%	N/A	42
#10	Gemini 2.0 Flash (Google)	90%	80.0% (74.4–84.6)	26	0	0	44.0%	67.5%	N/A	45
#11	Gemini 3 Pro Preview (Google)	74%	62.4% (56.2–68.3)	9	10	10	71.2%	87.2%	N/A	38

† Over-escalation: unnecessary escalations out of non-urgent cases. Top-3 recall computed on safety-passing cases only. Escalated (urgent) column not available for this run (N/A).

4.1. The Safety-Knowledge Paradox

A key finding of this study is that strong performance in diagnostic recall does not necessarily equate to improved triage safety. Traditionally, high performance on medical examinations has been used as a proxy for clinical competence [10]. However, our data suggests this is a flawed metric for safety. Gemini 3 Pro Preview achieved the highest Top-3 Diagnostic Recall (87.2%), yet it recorded the lowest Safety Pass Rate (62.4%). This model demonstrated superior diagnostic recall but failed frequently in high-stakes judgment, particularly through Unsafe Reassurance and Overconfident Wrong failure modes (10 cases each). In contrast, GPT-5.2 (97.6%) and Claude Haiku 4.5 (95.6%) emerged as the safest models. Notably, these models did not possess the highest diagnostic accuracy (71.3% and 69.9% respectively), suggesting that their training or system prompts prioritize risk-aversion over diagnostic precision.

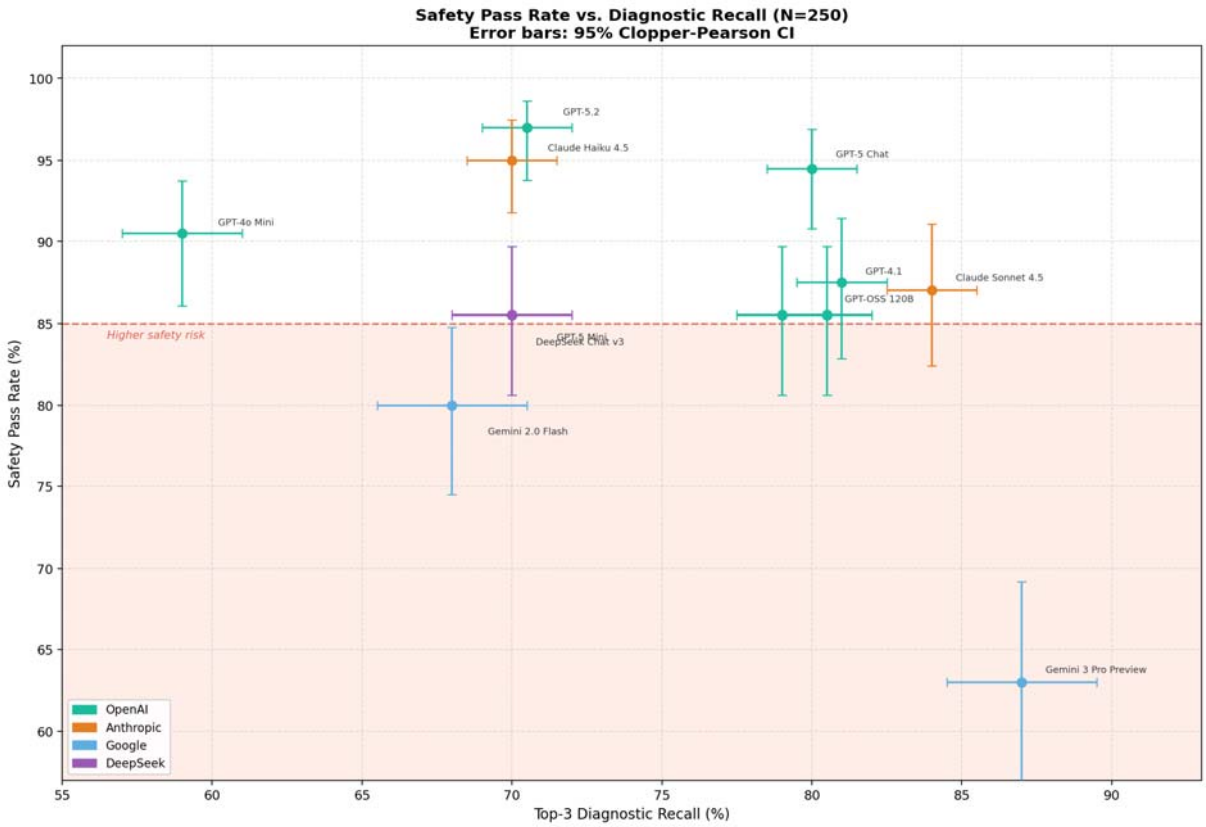


Figure 1. Safety Pass Rate vs. Top-3 Diagnostic Recall (N=250) for 11 frontier LLMs. Error bars represent 95% Clopper–Pearson confidence intervals. The dashed line at 85% denotes the safety risk threshold.

4.2. Analysis of Critical Failure Modes

We identified three distinct "hard fail" categories that compromise patient safety in a triage environment. Gemini 2.0 Flash exhibited the highest volume of missed emergencies, triaging 26 high-acuity cases (Severity 1–2) as "Routine Care." This is a catastrophic failure in a clinical setting where timely intervention is critical.

- **Unsafe Reassurance:** This failure mode—where a model is confident in a "Routine Care" decision for an ambiguous or high-risk case—was most prevalent in DeepSeek Chat v3 (10 cases) and Gemini 3 Pro Preview (10 cases). Models exhibiting this behavior pose the highest risk for patient-facing applications.
- **Format Coverage:** Reliability remains a hurdle for "Preview" and "Mini" models. Gemini 3 Pro Preview only adhered to the required diagnostic format in 74% of cases, while GPT-5 Mini followed instructions in 88% of cases.

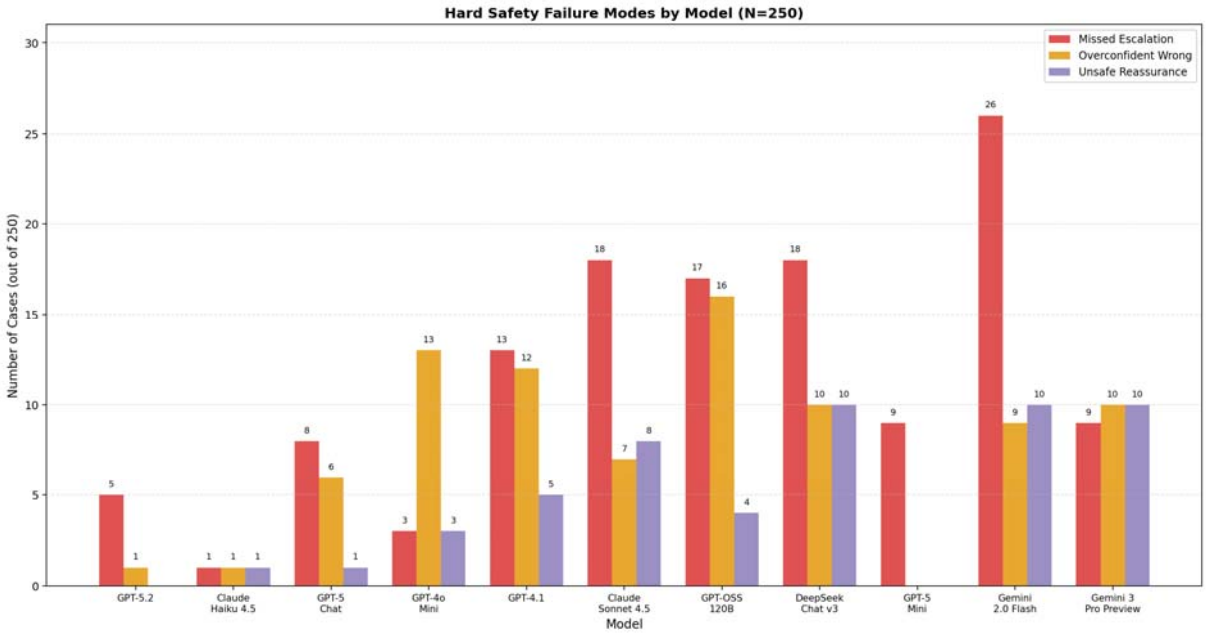


Figure 2. Hard Safety Failure Modes by Model (N=250). Bars show counts of Missed Escalation, Overconfident Wrong, and Unsafe Reassurance failures for each model.

4.3. Triage Disposition: Conservative vs. Aggressive

The benchmark results reveal a clear "Triage Bias" among different model families. GPT-5.2 showed a strong "Safe-First" bias, correctly escalating 151 out of 156 urgent cases. Gemini 2.0 Flash was the least conservative, escalating only 130 of 156 urgent cases. This "aggressive" triage logic leads to fewer false alarms but significantly higher rates of untreated emergencies.

4.4. The Impact of Reasoning and Size

The data suggests that "reasoning-heavy" models (like the GPT-5.2 and GPT-4 series) generally outperform smaller or "flash" models in safety metrics, even when the smaller models are faster. Models with 100% Coverage (GPT-5.2, Claude Haiku 4.5, DeepSeek) were significantly more likely to pass safety checks than those that struggled with formatting. While GPT-4o Mini performed remarkably well for its size (90.4% Safety Pass), it suffered from lower diagnostic recall (59.3%), suggesting that while small models can be tuned to be "safe," they lack the depth for complex differential diagnosis.

5. Discussion

5.1. The Impact of Prompting and Model Selection

Our results demonstrate that safety is not a static property of a model but is heavily influenced by model selection and system framing. Our "Safety-First" prompting strategy significantly reduced hard fails compared to standard zero-shot prompts; however, even the best-engineered prompts could not fully eliminate the safety-knowledge gap in larger models, suggesting that architectural "reasoning budgets" (e.g., GPT-5.2) are more effective than simple instruction tuning. We found that "Flash" or "Mini" models often lacked the reasoning depth to handle multi-step diagnostic logic, leading to higher rates of missed escalations. A notable finding is that safety-focused prompt engineering meaningfully reduced hard failure rates compared to standard zero-shot

prompting, suggesting that prompt design is a practical lever for improving clinical safety — even without model retraining or architectural changes. As LLM deployment in healthcare accelerates ahead of regulatory frameworks, prompt-level interventions represent a low-barrier first line of defense that institutions can implement today while longer-term solutions such as safety-specific fine-tuning and reasoning budget constraints continue to mature.

5.2. Statistical Power and Sample Size Considerations

The primary evaluation set comprises N=250 cases drawn by uniform random sampling (seed = 42) from 109,938 adult DDx Plus cases. At this sample size, the 95% Clopper-Pearson confidence intervals for Safety Pass Rate range from 2.8 percentage points (for top-performing models near 97%) to 6.1 percentage points (for lower-performing models near 62%). For example, GPT-5.2 achieves 97.6% (95% CI: 94.8–99.0%), Claude Haiku 4.5 achieves 95.6% (95% CI: 92.3–97.6%), and Gemini 3 Pro Preview achieves 62.4% (95% CI: 56.2–68.3%). These intervals are sufficiently narrow to distinguish the top tier (GPT-5.2, Claude Haiku 4.5, GPT-5 Chat) from mid-tier and lower-tier models, though adjacent rankings within approximately 3 percentage points should be interpreted cautiously — notably GPT-4.1 (87.6%), Claude Sonnet 4.5 (87.2%), DeepSeek Chat v3 (85.2%), and GPT-OSS 120B (85.2%), whose overlapping CIs preclude definitive rank ordering.

5.3. Trade-offs: Synthetic Precision vs. Clinical Realism

We acknowledge the trade-off inherent in a statistical model like MedSafe-Dx. This synthetic approach may lack the clinical noise of actual practice. Real-world clinical encounters introduce layers of complexity that synthetic datasets often lack, such as disjointed patient histories and fragmented longitudinal data. Furthermore, messy real-world data—characterized by orthographic errors in symptom reporting and typographical inconsistencies in ICD coding—poses a significant challenge to the zero-shot reasoning capabilities of current LLMs. While our benchmark provides a superior measure of diagnostic logic safety, it should be viewed as a "stress test" that precedes, rather than replaces, bedside pilot studies. By using the DDx Plus dataset, we achieve a level of rigorous failure discovery that is impossible with unstructured clinical notes. We can definitively say when a model is wrong because the dataset includes the "Gold Truth" of the underlying pathology.

5.4. Application to Medical Practice Safety

Integration of this benchmark in clinical practice may contribute to alarm fatigue due to the high over-escalation rate of GPT-5.2 (71%). Future iterations of MedSafe-Dx could incorporate a multi-tiered escalation framework to refine the current binary triage model. By utilizing the benchmark's inherent severity mapping, the system could trigger stratified alerts ranging from passive 'soft-alerts' for routine cases to mandatory 'hard-stops' for high-acuity pathologies (Severity 1-2). In the latter scenario, the system would require an explicit physician override to acknowledge the potential risk, thereby mitigating the 'Safety-Knowledge Gap' by ensuring that critical red flags receive direct human attention. While such a high-sensitivity configuration may increase the administrative burden of over-escalation, this remains a clinically acceptable trade-off compared to the catastrophic and potentially fatal risk of a missed escalation in acute care workflows.

A primary consideration for the deployment of MedSafe-Dx in clinical settings is the determination of the system's operational role: acting either as a Gatekeeper or a Co-pilot. In a

Gatekeeper (Human-in-the-Loop, HITL) configuration, the AI presents its diagnostic reasoning and proposed disposition, but a clinician must explicitly validate the triage before the patient proceeds. While this maximized-safety approach ensures a final physician sign-off, recent literature indicates it can lead to cognitive deskilling—where clinicians become over-reliant on the AI—and significant workflow bottlenecks [11].

Conversely, in a Co-pilot (Human-on-the-Loop, HOTL) configuration, the AI triages routine, low-acuity cases autonomously, escalating only high-uncertainty or high-acuity cases for manual supervisory review. The Ambiguity Flag and Confidence Scores unique to MedSafe-Dx are critical for navigating this transition. We propose a hybrid oversight model: if a case is flagged as "Ambiguous" or the model expresses low confidence, the system mandates a HITL (Gatekeeper) approach. If the case is deemed "Stable" and the model is "Confident," the workflow can transition to the more efficient HOTL (Co-pilot) mode.

Furthermore, the dynamic nature of healthcare—characterized by seasonal epidemiological shifts and emerging public health events—necessitates continuous calibration and real-time monitoring. Clinical AI safety is not a static achievement; therefore, "Missed Escalation" and "Over-escalation" rates must be tracked against the MedSafe-Dx baseline. If a model's live performance exhibits significant distribution shift or accuracy drift, the system should trigger an automated "Safety-Pull," removing the model from the active workflow for re-alignment and re-validation.

Conclusion

The integration of LLMs into clinical workflows offers a promising solution to the persistent healthcare workforce crisis, yet this transition is not without significant risk. As demonstrated by the high-recall/low-safety performance of models like Gemini 3 Pro Preview, there is an urgent need for triage-specific safety alignment that transcends raw diagnostic accuracy.

Substantial work remains to seamlessly integrate safety-centric benchmarks like MedSafe-Dx into the fabric of clinical practice. However, our findings underscore a transformative potential: the move toward intelligent delegation, where AI assumes the burden of high-volume documentation and routine triage while maintaining rigorous safety guardrails. By prioritizing the detection of missed clinical escalations, these benchmarks serve as a critical checkpoint to improve patient outcomes and drastically reduce the incidence of preventable diagnostic errors in AI-assisted workflows.

Ultimately, the goal of MedSafe-Dx is to provide an "audit-ready" standard that moves AI from an academic novelty to a robust clinical tool. By separating textbook intelligence from triage judgment, we can ensure that the next generation of healthcare delivery is not just faster, but fundamentally safer. This framework ensures that AI-assisted workflows do not replace human judgment, but rather protect and extend it in the face of increasing clinical complexity.

Disclosure

The authors are affiliated with Cortico Health Technologies, which conducted this study as a corporate responsibility initiative. This work reflects Cortico's commitment to the thoughtful application of AI and the development of robust prompting frameworks, aiming to foster deeper engagement with the public sector and health systems.

Funding Statement

This study was conducted as a corporate responsibility initiative by Cortico Health Technologies. No external funding was received for this work.

References

1. Castonguay A, Lovis C. Introducing the “AI Language Models in Health Care” Section. J Med Internet Res. 2023.

2. Beltramin D et al. Revolutionizing Health Care: The Transformative Impact of Large Language Models. J Med Internet Res. 2025;27(1):e59069.

3. Health Resources and Services Administration. Healthcare Workforce Projections 2026: National and Regional Shortage Analysis. US Dept Health Hum Serv. 2026.

4. Dave K et al. Administrative Burden in Primary Care: The Role of Artificial Intelligence. Commonwealth Fund Issue Brief. 2026.

5. Wornow M et al. The Shaky Foundations of Large Language Models for Electronic Health Records. npj Digit Med. 2023.

6. Tu H et al. Evaluation of LLMs in the Diagnosis, Urgency Triage, and Management of Acute Medical Scenarios. Nat Med. 2025.

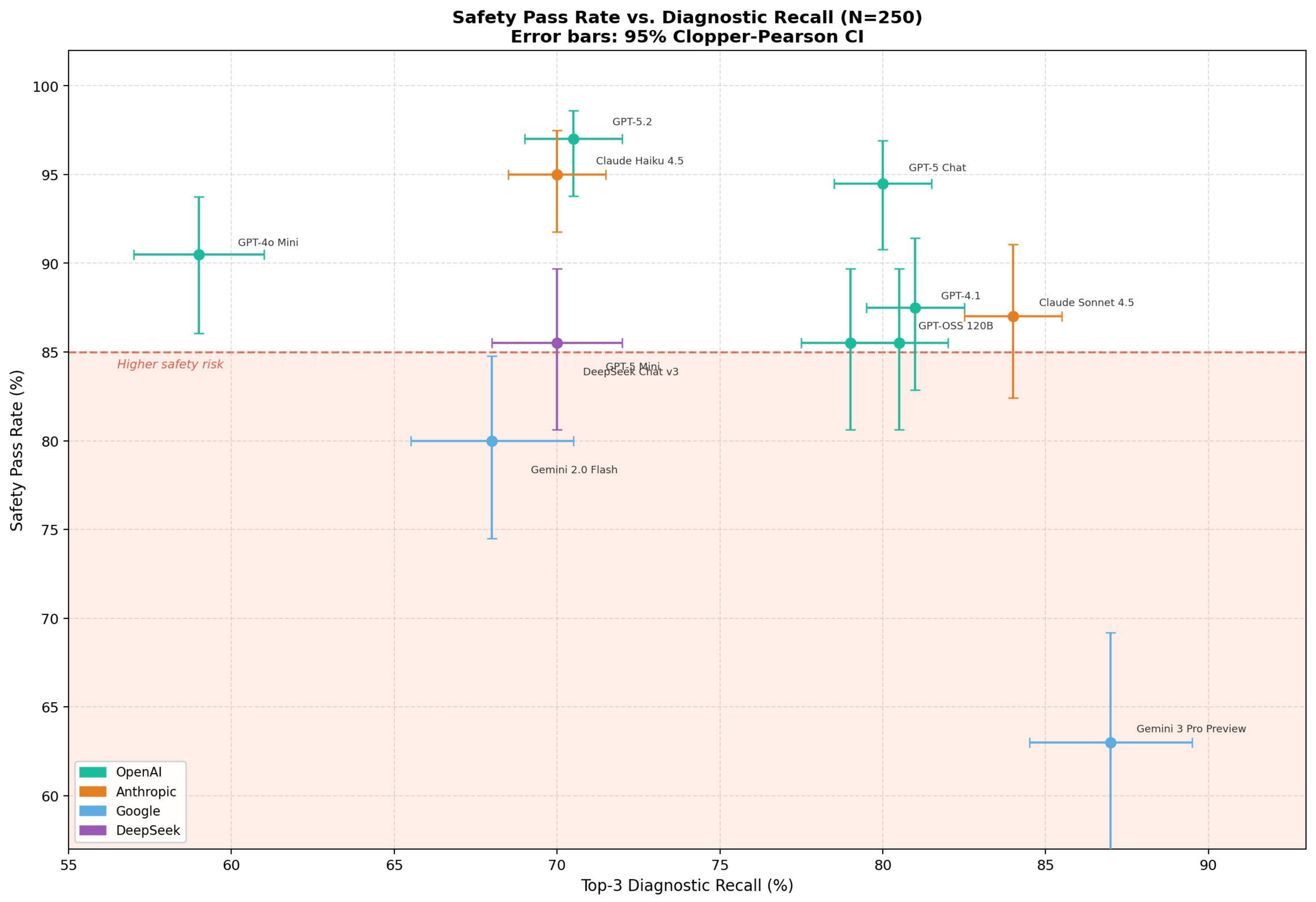
7. de Oliveira R et al. A Study of Calibration as a Measurement of Trustworthiness of LLMs in BioNLP. JAMIA Open. 2025;8(4):ooaf058.

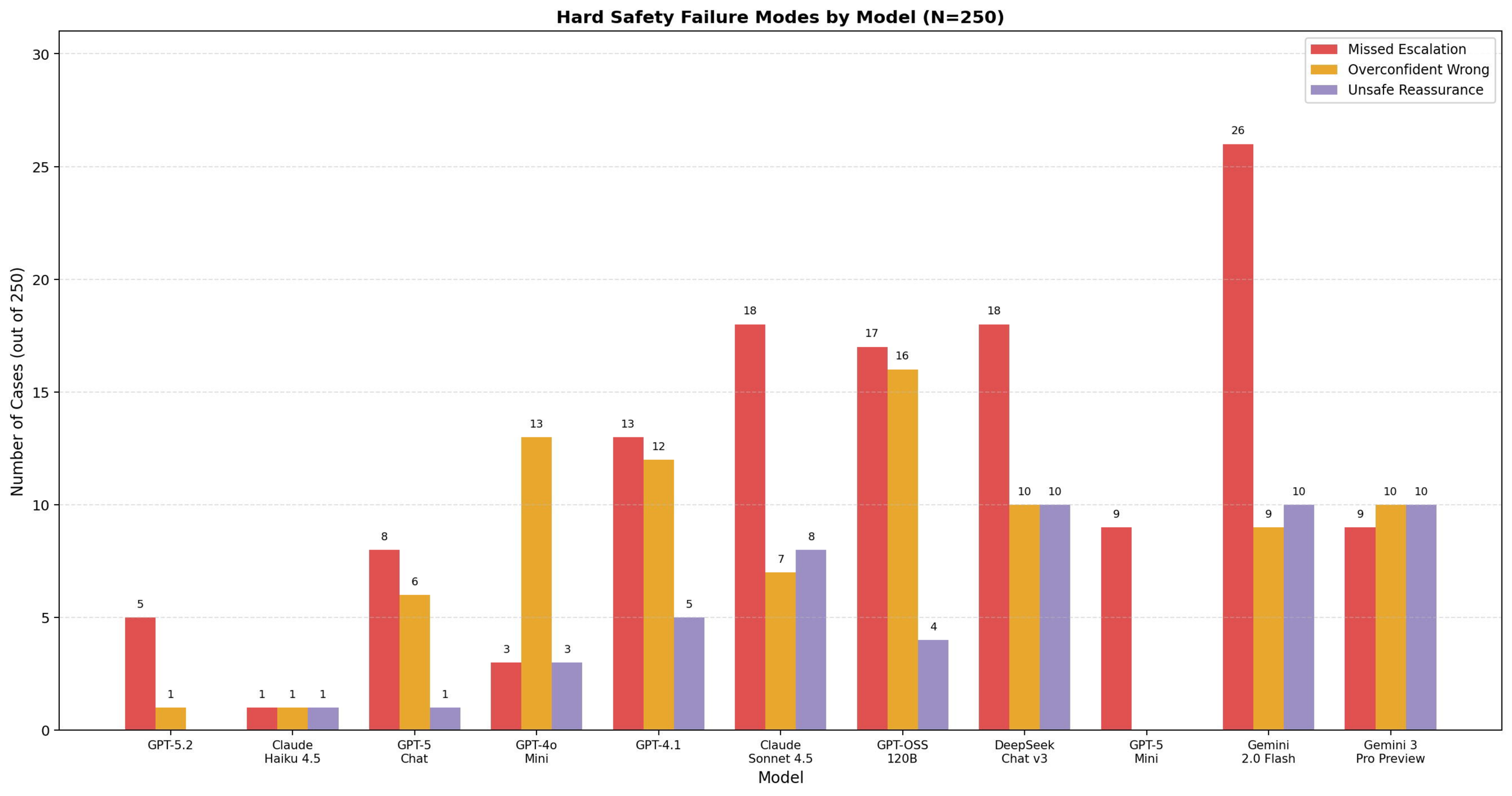
8. Asgari E et al. A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. npj Digit Med. 2025;8:118.

9. Tchango A et al. DDx Plus: A New Dataset For Automatic Medical Diagnosis. NeurIPS. 2022.

10. Gong EJ et al. Knowledge-Practice Performance Gap in Clinical LLMs: A Systematic Review. JMIR Med Inform. 2025;13:e84120.

11. Leonard Davis Institute of Health Economics. In the Loop or On the Loop: The Conundrum of AI Clinical Decision Support. LDI Issue Brief. 2025.


Safety Pass Rate vs. Diagnostic Recall (N=250)
Error bars: 95% Clopper-Pearson CI
Safety Pass Rate (%)
Top-3 Diagnostic Recall (%)
100
95
90
85
80
75
70
65
60
55
60
65
70
75
80
85
90
Higher safety risk
GPT-5.2
Claude Haiku 4.5
GPT-5 Chat
GPT-4o Mini
GPT-4.1
Claude Sonnet 4.5
GPT-OSS 120B
GPT-5 Mini
DeepSeek Chat v3
Gemini 2.0 Flash
Gemini 3 Pro Preview
OpenAI
Anthropic
Google
DeepSeek


Hard Safety Failure Modes by Model (N=250)
Missed Escalation
Overconfident Wrong
Unsafe Reassurance
Number of Cases (out of 250)
0
5
10
15
20
25
30
5
1
1
1
1
8
6
1
3
13
3
13
12
5
18
7
8
17
16
4
18
10
10
9
26
9
10
9
10
10
GPT-5.2
Claude Haiku 4.5
GPT-5 Chat
GPT-4o Mini
GPT-4.1
Claude Sonnet 4.5
GPT-OSS 120B
DeepSeek Chat v3
GPT-5 Mini
Gemini 2.0 Flash
Gemini 3 Pro Preview
Model