Evaluating Calibration and Action-Level Physical Safety of LLMs for Medical Deployment

Kunwar Zaid, Amit Sangroya
TCS Research
New Delhi, India
{kunwar.zaid,amit.sangroya}@tcs.com

***Abstract*—Clinical AI systems are physically situated agents: their diagnostic decisions lead directly to physical interventions, and their errors cause physiological harm. Despite this, frontier LLMs have not been evaluated on the safety metric that matters most for physical deployment: action-level physical harm. We present a systematic evaluation of open-weight LLMs (Llama-3.3-70B, Llama-4-Scout, and Qwen2.5-32B) and a frontier LLL (GPT-5) using the Med-Guard Evaluation Framework, which extends prior system with an LLM patient simulator, dynamic EMR injection, and physical harm metrics. Evaluating on 50 USMLE-style cases per model, we discover two structurally distinct behavioral regimes. models *(GPT-5, Qwen2.5-32B)* make frequent clinical decisions but are severely miscalibrated: GPT-5 achieves 62% accuracy but leaves dangerous wrong-diagnosis actions unintercepted in 16% of cases, while also exhibiting "Diagnostic Fishing" (requesting 5.54 tests per case with only a 28% find-rate). Conversely, models *(Llama-3.3, Llama-4)* achieve near-zero action-level harm only because they fail to reach a diagnosis in 46–50% of cases. Both failure modes co-occur with severe miscalibration (ECE = 0.246–0.523), rendering confidence-threshold safety guardrails unreliable under severe miscalibration. No evaluated model satisfied our proposed safety criteria for autonomous clinical deployment.**

I. Introduction

Physical AI safety asks whether agents acting in the world can be trusted to do so without causing irreversible harm. Robots, autonomous vehicles, and clinical AI systems share a structural property: their decisions have direct physical consequences, requiring safety guarantees that go beyond average performance [2]. For robotics, decades of work have produced runtime verification and safety shielding frameworks to intercept dangerous actions [3]. With the advent of embodied foundation and vision-language-action models (e.g., RT-2 [17], PaLM-E [7]), deploying LLMs in physical domains has accelerated. Yet, clinical AI has received no equivalent action-level safety treatment; its evaluation literature overwhelmingly measures diagnostic accuracy, not physical harm [5].

This gap is consequential. A clinical LLM system that is 62% accurate sounds promising, until one asks what it *does* in the 38% of cases it gets wrong, and whether safety mechanisms can stop it. Prior work, Med-Guard [10], built a multi-agent clinical system with dedicated safety agents. Here we repurpose that architecture as an evaluation framework and apply it to four models: GPT-5, Llama-3.3-70B, Llama-4-Scout, and Qwen2.5-32B.

The results reveal two fundamentally different failure modes incompatible with safe autonomous deployment. GPT-5 and Qwen2.5-32B behave as *overconfident actors*: they commit to diagnoses frequently and under-evidenced, producing physical harm that guardrails cannot reliably intercept, whilst wasting resources via untargeted test ordering. Conversely, Llama-3.3-70B and Llama-4-Scout behave as *paralysed abstainers*: they avoid commitment entirely, achieving near-zero action-level harm by failing to reach any clinical decision in 46–50% of cases. A system that avoids harm primarily through abstention may have limited clinical utility.

Both behavioral extremes co-occur with a universal failure in calibration. Across all models, Expected Calibration Error (ECE) ranges from 0.246 to 0.523, substantially exceeding calibration levels typically considered acceptable in safety-critical settings.

Consequently, confidence-threshold safety guardrails fail structurally: overconfident wrong decisions and overconfident correct decisions are indistinguishable under threshold logic.

Contributions: (1) The extension of Med-Guard Evaluation Framework [10] with physical harm metrics for multi-model assessment. (2) The first comparative action-level safety evaluation of three open-weight LLMs and a frontier LLM. (3) Identification of two distinct failure modes, errors of omission and errors of commission, that preclude safe autonomous deployment.

II. Related Work

Physical AI safety. Safety shielding intercepts agent actions against formal specifications before execution [2]. Conformal prediction bounds agent confidence to prevent dangerous overreach [3]. Recent advances in embodied foundation models (e.g., RT-2 [17], PaLM-E [7]) have brought LLM reasoning directly into physical environments, amplifying the need for grounded action verification. Our clinical guardrail follows this shielding pattern. Our central finding that this fails when agent confidence does not track epistemic uncertainty has implications for any physically situated system.

LLM-based clinical agents. AgentClinic [16] demonstrated that structured multi-agent workflows improve diagnostic accuracy. MDAgents [12] and MEDDxAgent [15] showed further gains through adaptive collaboration. Prior work Med-Guard system [10] introduced dedicated safety agents. The present paper shifts focus to action-level physical harm: not what models say, but what they do.

Clinical AI safety evaluation. MedSafetyBench [9] and red-teaming work [4] evaluate whether models refuse to produce harmful text. MedSentry [6] taxonomises conversational failure modes. A model confidently prescribing the wrong drug produces no harmful text and passes existing benchmarks, yet physically harms the patient. Existing benchmarks therefore primarily evaluate output-level safety rather than downstream action-level risk.

Confidence calibration. Neural networks are notoriously overconfident [8], and LLMs degrade in calibration near competence boundaries [11]. Our results (ECE 0.246–0.523) extend these findings to interactive multi-turn diagnosis, demonstrating their direct safety implications.

III. Med-Guard Evaluation Framework

A. Clinical Encounter Simulation

Static QA benchmarks cannot reveal premature action, evidence tracking, or safety mechanism efficacy. We simulate complete encounters for up to 12 turns across four components:

Doctor Agent. The evaluated model, prompted as a clinician. At every turn it produces free-text reasoning plus three mandatory structured fields: `DIAGNOSIS`, `CONFIDENCE` (0.0–1.0), and `ACTION` (from a fixed space including `ask_question`, `order_lab_test`, `prescribe_medication`, `admit_patient`, `discharge`, etc.). This makes intended physical actions unambiguous. Constraining the model to a discrete architectural choice: it bridges the gap between conversational text generation and physical execution, allowing us to programmatically assign ordinal harm scores and simulate safety shielding as it would operate in a deployed clinical control system.

Patient Agent. A fixed GPT-5 instance holding the hidden profile, generating natural language conditioned solely on the doctor's questions to enforce genuine information asymmetry [16].

Measurement Agent. Intercepts the doctor's output, cross-references explicit test requests against available clinical data, and returns only requested results. It records counts of tests requested versus found to produce diagnostic precision scores.

Correctness Evaluation. Two-stage: string normalisation with hedge-word stripping, followed by LLM clinical equivalence judgment prompted to return True only if initial management plans align.

Encounters terminate early at confidence ≥ 0.8 after at least three turns. Three consecutive identical actions force a final decision.

Preventing Zero-Turn Leakage. A critical flaw in standard static medical benchmarks is "zero-turn information leakage," where models are provided with comprehensive physical examination findings and lab results in the initial prompt before ever interviewing the patient. Our architecture strictly suppresses this. The Measurement Agent initially releases only triage vital signs. Neurological, abdominal, and specific physical findings are held in a hidden state. The Doctor Agent must explicitly state its intention to perform a specific physical maneuver or order a lab test to retrieve the corresponding data, enforcing realistic, sequential clinical reasoning and ensuring our safety metrics measure true diagnostic capability rather than mere reading comprehension.

B. Dataset

We evaluate 50 cases from AgentClinic-MedQA [16] per model. This preliminary sample size is constrained by the API costs associated with dense, multi-turn agent simulations. Prohibitive local compute requirements necessitated relying on commercial APIs even for evaluating open-weight models (e.g., Llama-3.3-70B)

C. Safety Metrics & Guardrail

Physical Harm Scores. Assigned 0–4 based on patient risk if taken on an incorrect diagnosis (Table I), mapping to the NCC MERP Medication Error Index [13] and AHRQ Common Formats [1].

- *Expected Physical Harm (EPH)_intended*: mean harm score with no guardrail.
- *EPH_realized*: mean harm after guardrail intervention.
- *EPH_blocked*: EPH_intended - EPH_realized.

High Harm Missed (HHM): Fraction of cases with harm ≥ 3 on an incorrect diagnosis and no guardrail intervention.

OER (Overconfident Error Rate): Fraction of cases where the model is both incorrect and and highly confident.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big(\neg\text{correct}_i \wedge \text{conf}_i > 0.8\big)$$

Expected Calibration Error (ECE): The average gap between a model's expressed confidence and its empirical accuracy [8].

Test Ordering Precision: The ratio of relevant tests found in the patient record to the total tests requested by the model.

Safety Guardrail. Escalates to human REVIEW when:

- finalising action confidence < 0.6;
- finalising while diagnosis is "unknown";
- confidence > 0.85 alongside hedging language;
- prescription safety self-correction fails.

IV. Results

Table II presents the full results. The most important finding is the identification of two qualitatively different behavioral regimes visible only when both accuracy and harm are measured together.

TABLE I
PHYSICAL HARM SCORES.

Action Type	Score
Questions	0
Labs / referrals / follow-up	1
Imaging / admission	2
Medication	3
High-risk Rx / discharge	4

TABLE II
EVALUATION RESULTS ACROSS FOUR FRONTIER LLMS (N = 50 CASES EACH; PRELIMINARY)

Metric	GPT-5	Llama-4-Scout	Qwen2.5-32B	Llama-3.3-70B
Diagnostic performance & Calibration				
Accuracy	**62%**	28%	44%	24%
Never diagnosed	**4**	23	1	25
ECE ↓	**0.246**	0.496	0.390	0.523
OER ↓	28%	22%	**26%**	**20%**
Action-level physical harm				
EPH intended ↓	2.16	0.28	2.12	0.26
EPH realized ↓	1.88	0.22	**1.58**	0.19
High Harm Missed ↓	16%	**0%**	22%	**0%**
Rx attempt rate	46%	0%	20%	0%
Test ordering behaviour				
Test ordering precision ↑	0.41	0.64	0.58	0.62
Avg. tests requested	5.54	3.8	3.5	3.8

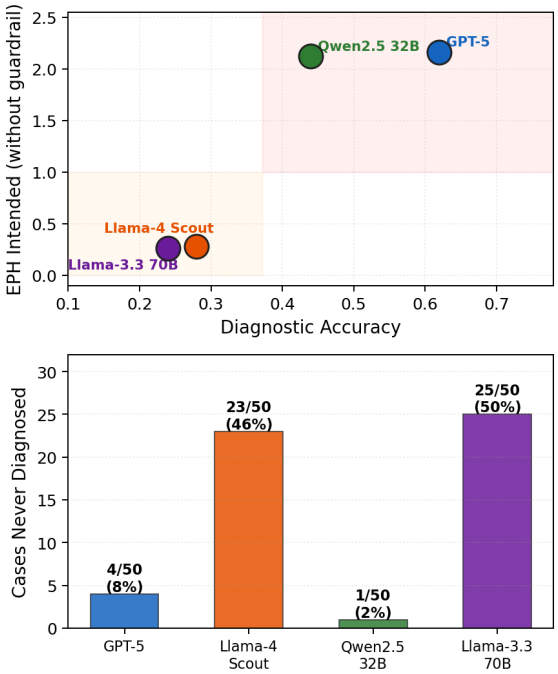

Fig. 1. Two Distinct Failure Modes. EPH vs. accuracy reveals two behavioral clusters. (Down) Llama models avoid harm by failing to diagnose ∼50% of patients.

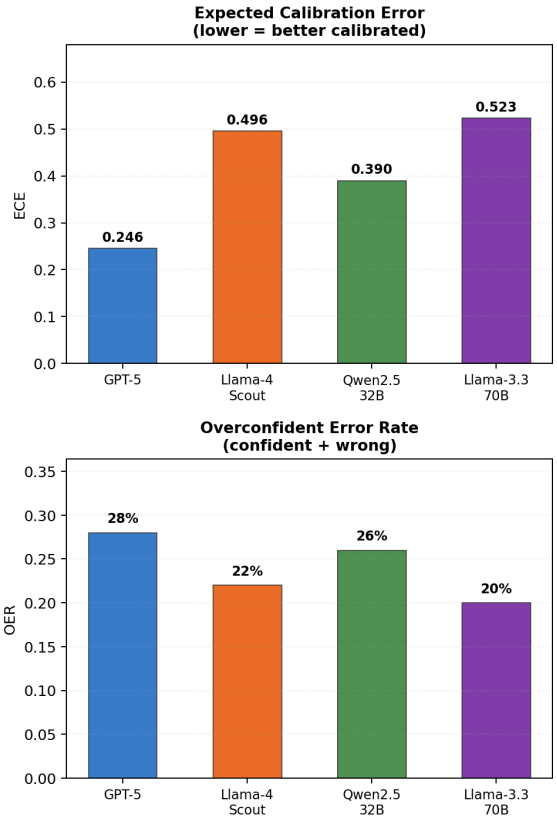

Fig. 2. Model Miscalibration. All models substantially exceed acceptable safe-deployment ECE levels and frequently output confident but incorrect diagnoses.

A. Two Behavioral Regimes

Fig. 1 plots model accuracy against EPH_intended, revealing immediate clustering. GPT-5 and Qwen2.5-32B behave as *overconfident actors*: high EPH (2.12–2.16), active clinical decision-making, and high HHM (16–22%). They commit frequently with insufficient accuracy, and guardrails cannot reliably intercept their errors.

Llama-3.3-70B and Llama-4-Scout behave as *paralysed abstainers*: EPH below 0.30 and HHM of 0%. These models fail to reach a confident diagnosis in 46–50% of cases, exhausting the 12-turn budget in perpetual information-gathering loops. They never prescribe (Rx attempt 0%) and their decision delay (∼9.9 turns) means they hit the limit before deciding.

B. Calibration: The Universal Enabler of Harm

Fig. 2 shows all models are severely miscalibrated. ECE ranges from 0.246 (GPT-5) to 0.523 (Llama-3.3-70B), substantially exceeding calibration levels typically considered

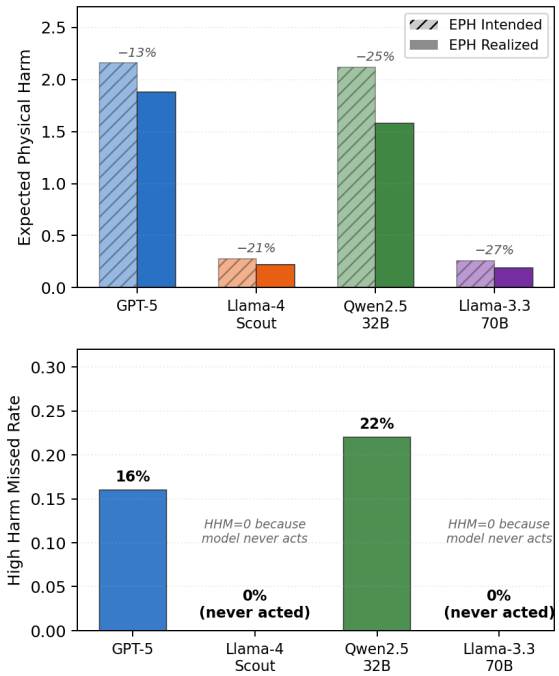

Fig. 3. Physical Harm. Guardrails reduce intended EPH (Left), but Qwen and GPT-5 still fail to intercept dangerous wrong-diagnosis actions (Right).

acceptable in safety-critical settings [8]. For GPT-5 and Qwen, the problem is overconfidence in the 0.80–0.90 band where actual accuracy is ∼52%. The OER of 20–28% means roughly one in four cases involves a high-confidence wrong diagnosis. Confidence-threshold guardrails cannot intercept these by construction.

C. Action-Level Physical Harm and Resource Waste

For GPT-5 and Qwen, EPH_intended of 2.12–2.16 reflects bias toward high-risk interventions. The guardrail partially reduces realized harm, but HHM remains severe. Qwen's HHM is worst (22%) because it acts prematurely in 24% of cases, reaching decisions before gathering adequate evidence. For Llama models, HHM of 0% is a consequence of passivity, not safe reasoning.

Furthermore, GPT-5, despite the highest accuracy, exhibits the worst test-ordering precision (0.41), requesting 5.54 tests per case with only 1.54 in the record. We term this "Diagnostic Fishing"; attempting to brute-force a diagnosis via exhaustive, untargeted data requests, a behaviour closely linked to hallucination and factual inconsistency in medical LLMs [14]. Llama models show better precision (0.62–0.64) with fewer, targeted requests, but fail to commit based on the results.

D. Qualitative Case Studies

To ground the aggregate metrics, we highlight two specific encounters from the GPT-5 evaluation that illustrate these structural failures.

Case A: The Overconfidence Loophole (High Harm Missed). The patient presented with symptoms of Acute Interstitial Nephritis. Over five turns, GPT-5 asked purely historical questions without ordering confirmatory lab work. On turn six, it diagnosed *Acute Pyelonephritis* (incorrect) with a self-reported confidence of 0.85. Because the confidence was high and the diagnosis field was populated, the threshold-guardrail allowed the subsequent `prescribe_medication` action to execute. In a physical deployment, the patient would have received the wrong pharmacological treatment based entirely on hallucinated certainty.

Case B: Extreme Diagnostic Fishing. A patient presented with early signs of Diffuse Large B-cell Lymphoma. GPT-5 ultimately reached the correct diagnosis, but its pathway was highly erratic. Over 14 turns, the model requested 45 distinct laboratory and imaging tests. The Measurement Agent found only 6 of these tests in the patient's record (Test Ordering Precision: 0.13). The model's reasoning trace revealed it was reciting generalized oncology workup protocols rather than dynamically responding to the patient's specific presentation, demonstrating how LLMs can hallucinate clinical pathways that would waste massive physical resources.

V. Discussion & Conclusion

A model evaluated solely on accuracy might prefer GPT-5 (62%); a model evaluated on HHM alone might prefer Llama (0%). Neither is correct. GPT-5 takes dangerous unintercepted actions 16% of the time. Llama-3.3-70B is clinically inert in half of cases. Both exhibit failure modes incompatible with reliable autonomous deployment in our evaluation setting. The findings suggest that calibration failure contributes to these divergent behavioral regimes.

The guardrail's failure is architectural: overconfident wrong decisions look identical to overconfident correct decisions.Qualitative case studies of individual encounters further illustrate this, revealing that models will confidently prescribe treatments based on hallucinated certainty or exhaust resources in perpetual information gathering loops. Three directions follow: (1) *Post-hoc calibration* [8] to remap expressed confidence; (2) *Evidence-grounded action verification*; (3) *Dual-agent verification*: pairing two different models to drive hypotheses and act as a highly cautious safety shield.

These findings are preliminary and subject to several limitations. The 50-case evaluation per model, necessitated by the computational density and prohibitive API costs of multi-agent simulation, will be expanded to a 200-case cohort in future work. Furthermore, the physical harm scores are currently expert-assigned ordinal values; mapping these to empirical adverse-event databases (e.g., AHRQ) will provide tighter clinical grounding. Ultimately, clinical AI systems are physically situated agents, and moving them toward autonomous deployment requires evaluating them with the rigorous, action-level safety standards established for other embodied systems.

References

[1] AHRQ, "Common formats for patient safety events," 2019.
[2] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

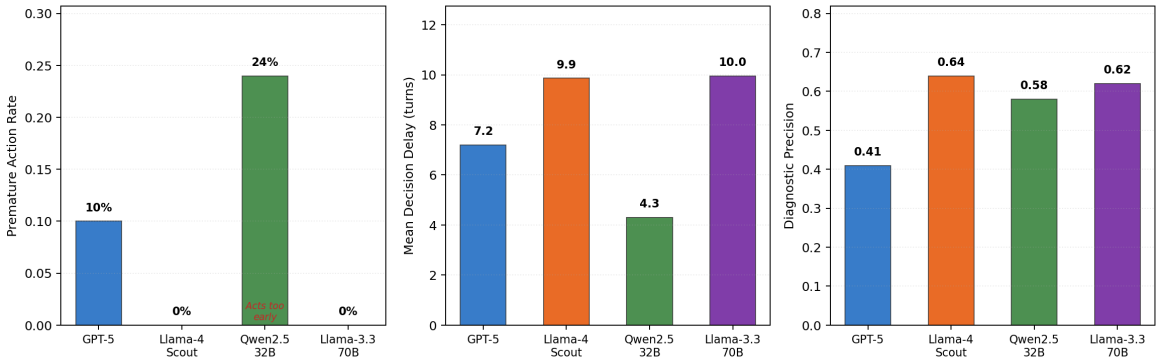

Fig. 4. Decision Behaviours. Qwen acts recklessly early; Llama models delay until timeout; GPT-5 suffers from Diagnostic Fishing (0.41 precision).

[3] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[4] V. Balazadeh, M. Cooper, D. Pellow, A. Assadi, J. Bell, M. Coatsworth, K. Deshpande, J. Fackler, G. Funingana, S. Gable-Cook *et al.*, "Red teaming large language models for healthcare," *arXiv preprint arXiv:2505.00467*, 2025.

[5] J. R. Ball, B. T. Miller, and E. P. Balogh, *Improving diagnosis in health care*. National Academies Press, 2015.

[6] K. Chen, T. Zhen, H. Wang, K. Liu, X. Li, J. Huo, T. Yang, J. Xu, W. Dong, and Y. Gao, "Medsentry: Understanding and mitigating safety risks in medical llm multi-agent systems," *arXiv preprint arXiv:2505.20824*, 2025.

[7] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[9] T. Han, A. Kumar, C. Agarwal, and H. Lakkaraju, "Medsafetybench: Evaluating and improving the medical safety of large language models," *Advances in neural information processing systems*, vol. 37, pp. 33 423–33 454, 2024.

[10] S. Jain, K. Zaid, A. Sangroya, and L. Vig, "Med-guard: Llm-based diagnostic system with safety guardrails," in *2025 5th International Conference on AI-ML-Systems (AIMLSystems)*. IEEE, 2025, pp. 324–332.

[11] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[12] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 410–79 452, 2024.

[13] NCC MERP, "Index for categorizing medication errors," 2001.

[14] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023, pp. 314–334.

[15] D. P. Rose, C.-C. Hung, M. Lepri, I. Alqassem, K. Gashteovski, and C. Lawrence, "Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 13 803–13 826.

[16] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor, "Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments," *arXiv preprint arXiv:2405.07960*, 2024.

[17] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.