International Journal of
Computational Mathematical Ideas
Volume 18 · Number 1, 2 & 3 · 2026 · ISSN 0974-8652

Available online at www.ijcmi.com

Design of Healthcare Agentic AI Systems: Medical Reasoning, Verification Gates, and Human Oversight

Setu S. M. Kodi [1,*] **Sudheer Singamsetty** [2]

[1] *Pre-Medical, University of the Incarnate Word, San Antonio, Texas, USA*
[2] *Data Management Consultant, EDNA Technology Consulting Limited,*
68 Forest Ridge Avenue, Waterdown, Ontario, Canada, L8B 1V4

[*] Corresponding author: setu.kodi@gmail.com

Article Info

Article history:
Received: 16-01-2026
Revised: 15-07-2026
Accepted: 26-07-2026
Available online: 2026

Keywords:
Healthcare AI; Agentic AI; Clinical decision support; Medical reasoning; Human-in-the-loop; Verification and guardrails; Patient safety.

Abstract

Agentic artificial intelligence, in which a model reasons, calls tools, and acts in a closed loop rather than emitting a single prediction, is arriving in medicine, where the tolerance for confident error is close to zero. This paper takes a design stance on healthcare agentic AI: rather than propose a new model, it asks how to build a clinical agentic system so that its autonomy is bounded by verification and human oversight. We distil six design requirements from the clinical AI literature, covering harm minimisation, human oversight, calibrated uncertainty, verification and provenance, equity, and interpretability. We give a reference architecture in which a reasoning agent is wrapped by a verification gate and a clinician escalation path, and we formalise that gate. We prove two guarantees: escalation is monotone in safety, so widening it never increases expected harm, and a risk-weighted gate minimises expected harm for any fixed amount of clinician effort. We validate the design with a fully reproducible simulation of an agentic clinical triage loop. The escalation gate trades autonomy for safety along a smooth curve, halving expected harm as oversight rises; layering an independent verifier and human review cuts the unsafe-action rate from 0.230 with the agent alone to 0.077; and risk-weighted escalation attains lower harm than confidence-only escalation at every clinician budget, exactly as the theory predicts. The results argue that in medicine the safety of an agentic system is engineered at the level of the loop, through verification, risk-weighted escalation, and clinician oversight, and can be designed and reasoned about rather than left to chance.

1. Introduction

Medicine is a high-stakes setting for automation. A recommendation that would be harmless in most software becomes a matter of safety when it concerns a patient, and the tolerance for confident error is close to zero. Over the past decade machine learning has nonetheless moved from the laboratory into the clinic, first through deep networks that match specialists on narrow perception tasks such as lesion classification [12] and retinal screening [13], and more recently through large language models that can carry on clinical conversations and answer medical questions at a level once thought distant [9]. The natural next step, and the subject of this paper, is the design of agentic systems: models that do not merely predict but observe, reason, call tools, and act inside a clinical workflow [5].

The promise is real and so is the danger. An agent that can retrieve a guideline, compute a dose, and place an order is far more useful than a passive classifier, but it can also act on a hallucination, propagate a hidden bias, or fail silently when the patient in front of it does not resemble its training data [21], a risk sharpened by the narrow scope of many clinical foundation models [11]. The lesson from the first wave of deployed clinical models is that raw accuracy is not the binding constraint; integration, oversight, and trust are [18]. A healthcare agentic system must therefore be engineered around its own fallibility.

This paper takes a design stance. Rather than propose a new model, we ask how to build a healthcare agentic AI system so that its autonomy is bounded by verification and human oversight, and we make the argument precise. Our contributions are three. First, we set out the design requirements that separate a safe clinical agent from a merely

capable one, drawing on the deployment literature [19] and its sociotechnical framing [20]. Second, we give a reference architecture in which a reasoning agent is wrapped by a verification gate and a clinician escalation path, and we formalise that gate: we prove that escalation is monotone in safety, so that widening it never increases expected harm, and that a risk-weighted gate minimises harm for any fixed amount of clinician effort. Third, we validate the design with a fully reproducible simulation of an agentic triage loop, measuring the trade-off between autonomy, safety, and human workload. All code and numbers are released so that every figure can be regenerated [1].

2. Background: Medical AI and the Turn to Agents

2.1. Deep learning reaches the clinic

The modern era of clinical AI began with perception. Convolutional networks reached dermatologist-level accuracy on skin lesion images [12] and matched ophthalmologists at grading diabetic retinopathy from fundus photographs [13]. The authors' own work continues this line, with deep models for brain tumour segmentation and survival prediction [2] and for early breast cancer classification [3]. Surveys of the period framed a convergence of human and machine capability across imaging, workflow, and patient-facing tools [14], while noting that translation into practice would be governed as much by trust and process as by performance [15].

2.2. Large language models and medical reasoning

Language models changed the scope of what clinical AI could attempt. Trained and tuned on medical corpora, they encode a broad base of clinical knowledge and can answer questions, summarise notes, and explain findings [9]. Reviews of the field catalogue both the opportunities and the failure modes, from hallucination to brittle reasoning [10]. A sober assessment of foundation models built on electronic health records warns that much of the reported promise rests on narrow datasets and evaluations that do not measure clinical usefulness [11]. Medical reasoning, in other words, is not yet a solved capability, and a system design must assume the model will sometimes be wrong.

2.3. Agentic methods

The agentic turn gives a model a loop. Interleaving reasoning traces with actions lets an agent gather evidence before committing [5]; teaching it to call external tools grounds its answers in calculators, databases, and guideline lookups rather than parametric memory alone [6]; and letting it reflect on its own failures improves subsequent attempts [7], and surveys now catalogue the resulting family of autonomous agents [8]. Reinforcement-style optimisation has produced agents that learn treatment policies directly from clinical data [16]. These mechanisms are powerful precisely because the agent acts, which is also why a clinical deployment must constrain what it is allowed to do on its own [4].

2.4. Lessons from real deployments

The clearest guidance comes from systems that actually reached the bedside. Early warning scores for sepsis were validated and then, more tellingly, integrated into routine care, where first validated [17] and then integrated into routine care, where the hard problems turned out to be workflow, alerting, and clinician trust rather than the model itself [18]. The decision-support literature had anticipated much of this: effective support fits the clinician's workflow, fires sparingly, and earns its place [19], and health information technology behaves as a sociotechnical system in which the tool, the people, and the organisation must align [20]. A deployed model also carries a lasting maintenance burden, the hidden technical debt of machine learning systems that accrues long after the first release [24]. A healthcare agentic system inherits all of these constraints and adds the new one that the tool now acts.

3. Design Requirements for Healthcare Agentic AI

From this literature we distil six requirements that a clinical agentic system must meet. They motivate the architecture of the next section.

Harm minimisation over accuracy. The objective is not to maximise average correctness but to minimise expected harm, which weights each error by the clinical cost of getting that case wrong [14]. A design that improves accuracy while occasionally acting catastrophically on a high-risk case is a bad design.

Human oversight by construction. A clinician must remain in the loop for consequential decisions, not as an afterthought but as a routed, resourced part of the system [20]. Oversight that depends on a busy clinician noticing a silent error is not oversight.

Calibrated uncertainty. The agent must know when it does not know. A gate can only escalate the right cases if the confidence signal it reads is meaningful, which places calibration at the centre of the design [11].

Verification and provenance. Proposed actions should be checked against explicit safety constraints and guideline sources before execution, and every action should carry an auditable trail [19]. Post hoc explanation is not a substitute

for a verifiable pipeline, a point argued forcefully in the medical context [22], where the role of explainability is itself contested across disciplines [23].

Equity. Clinical algorithms can encode and amplify bias when a convenient proxy stands in for the true target, as when cost is used as a proxy for need [21]. Fairness must be a design-time concern, monitored across subgroups.

Interpretability where it counts. For the highest-stakes decisions, an interpretable model or an inspectable decision path is preferable to a black box with a plausible story attached [25]. These requirements do not demand less capable agents; they demand agents embedded in a safer system.

4. Reference Architecture

Figure 1 gives a reference architecture that satisfies the six requirements. The forward path is a standard agentic pipeline: clinical inputs are gathered, relevant records and guidelines are retrieved and supplied to the model through retrieval-augmented generation, and a reasoning agent plans and calls tools to propose an action [5]. The two additions are what make the architecture clinical. A verification gate sits between the agent and any action: it checks the proposal against explicit safety constraints, reads the agent's calibrated confidence, and estimates the clinical risk of the case [6]. Depending on that assessment the gate either allows the action or escalates the case to a clinician, whose decision is authoritative. A monitoring and audit layer records every action and outcome, feeds a slow update loop, and watches for drift and subgroup disparities [21].

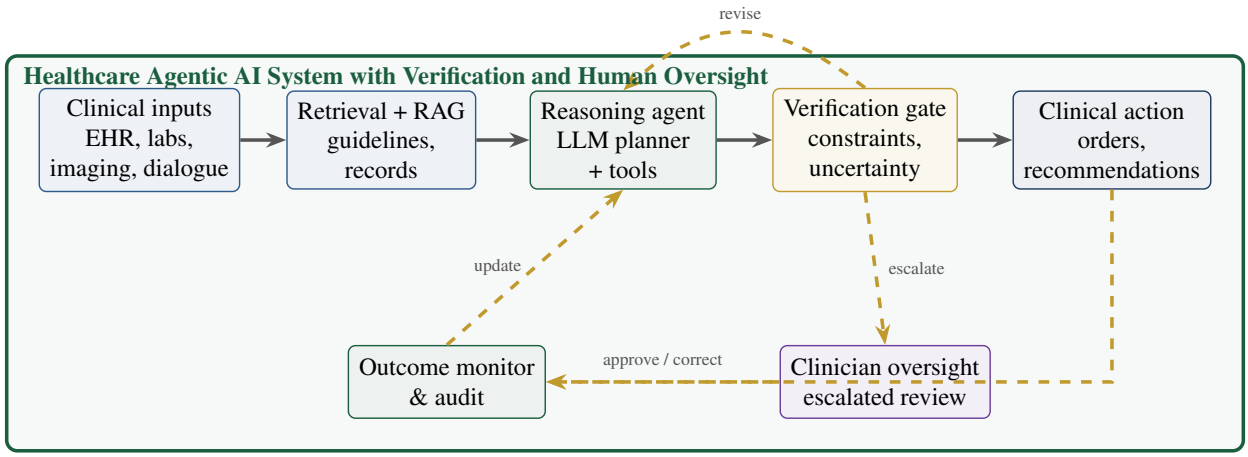

Figure 1. Reference architecture for a healthcare agentic AI system. A reasoning agent proposes actions on a retrieval-grounded forward path; a verification gate checks each proposal and escalates uncertain or high-risk cases to a clinician; a monitoring layer audits outcomes, updates the agent, and watches for drift and subgroup disparities.

5. Formal Model: Verification Gate and Escalation

We now formalise the gate and prove two guarantees. Index the stream of cases by i. Each case has a clinical risk $\rho_i \in [0,1]$, the harm incurred if its decision is wrong, an agent error probability e_i^{a}, and a human error probability e_i^{h}. Throughout we assume the clinician is at least as reliable as the agent on any case, $e_i^{\mathrm{h}} \le e_i^{\mathrm{a}}$, which is the reason a human is the escalation target [16]. The gate chooses an escalation set E; escalated cases are decided by the clinician and the rest by the agent. The expected harm is

$$H(E) = \sum_{i \notin E} \rho_i e_i^{\mathrm{a}} + \sum_{i \in E} \rho_i e_i^{\mathrm{h}}. \tag{1}$$

Writing $H_0 = \sum_i \rho_i e_i^{\mathrm{a}}$ for the no-escalation harm and $g_i = \rho_i (e_i^{\mathrm{a}} - e_i^{\mathrm{h}}) \ge 0$ for the harm reduction of escalating case i, equation (1) rearranges to

$$H(E) = H_0 - \sum_{i \in E} g_i. \tag{2}$$

Proposition 1 (Escalation is monotone in safety)**.** *If $E \subseteq E'$ then $H(E') \le H(E)$. Widening the escalation set never increases expected harm.*

Proof. From (2), $H(E) - H(E') = \sum_{i \in E' \setminus E} g_i$. Every term is a product of a nonnegative risk and a nonnegative error gap $e_i^{\mathrm{a}} - e_i^{\mathrm{h}} \ge 0$, so the sum is nonnegative and $H(E') \le H(E)$. □

Proposition 1 says the safety dial only turns one way: escalating more cases can cost clinician time but cannot make expected harm worse. The design question is therefore how to spend a limited escalation budget well.

Proposition 2 (Risk-weighted escalation is budget-optimal)**.** *Among all escalation sets with at most k cases, expected harm is minimised by escalating the k cases with the largest harm reduction g_i.*

Proof. By (2), minimising $H(E)$ subject to $|E| \leq k$ is equivalent to maximising $\sum_{i \in E} g_i$ subject to $|E| \leq k$. Since every $g_i \geq 0$, the sum is maximised by choosing the k largest values of g_i; any set omitting a larger g_i in favour of a smaller one can be improved by the swap, so the top-k selection is optimal. □

The optimal score $g_i = \rho_i(e_i^{\mathrm{a}} - e_i^{\mathrm{h}})$ multiplies risk by the agent's error gap, so a budget-optimal gate is risk-weighted, not confidence-only. A gate that escalates purely on low confidence approximates the error term but ignores ρ_i, and will spend its budget on low-risk cases the agent happens to find hard [22]. Algorithm 2 states the resulting clinical loop.

Algorithm 1: Agentic clinical decision loop with verification and escalation

Input: case; confidence threshold T; risk threshold; escalation budget.

1. retrieve records and guidelines (RAG); supply as context
2. agent proposes action a with reasoning trace and confidence c
3. verifier checks a against safety constraints; estimate risk ρ
4. compute escalation score $g = \rho\,(1 - c)$ as a harm-reduction proxy
5. **if** constraint violated **or** g above budget cutoff **then**
6. escalate to clinician; clinician decision is authoritative
7. **else** execute agent action a
8. record action, reasoning, and outcome to the audit log
9. monitor outcomes and subgroups; on drift, update the agent

Output: executed decision with provenance.

Figure 2. The clinical decision loop. The verifier and the risk-weighted escalation score (line 4) implement Propositions 1 and 2; the audit and monitoring steps (lines 8 to 9) close the oversight loop.

6. Simulation Study

To ground the design we simulate the loop from first principles in Python with a fixed seed, so every number is reproducible. We generate a population of cases, each with a difficulty, a clinical risk, an agent whose accuracy falls with difficulty, and a clinician who is more accurate, especially on hard cases. Outcomes are sampled and we measure accuracy, the unsafe-action rate (a wrong decision on a high-risk case), expected harm, and the clinician escalation rate. No external learning library is used.

6.1. The safety-autonomy trade-off

Figure 3 sweeps the confidence threshold at which the gate escalates. As the threshold rises the system hands more cases to the clinician: accuracy climbs from 0.77 toward 0.88, the unsafe-action rate falls from 0.22 to 0.11, and expected harm halves, at the cost of an escalation rate that grows to more than half of cases. The curve makes the design trade-off explicit and identifies an operating region in which most unsafe actions are removed at a moderate oversight cost [18].

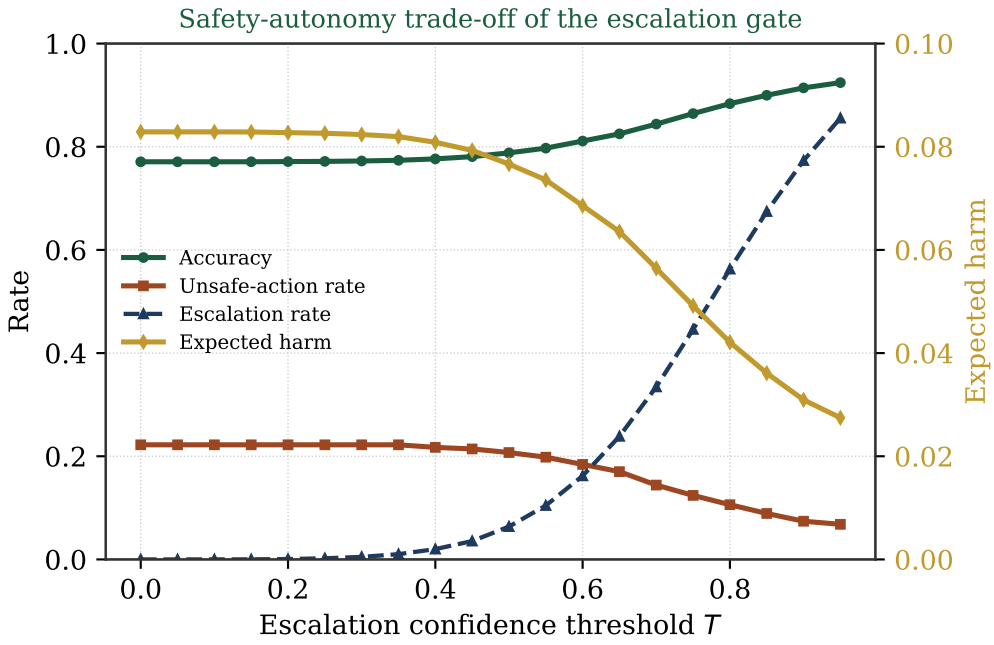

Figure 3. Safety-autonomy trade-off of the escalation gate. Raising the escalation threshold trades clinician workload for higher accuracy, fewer unsafe actions, and lower expected harm.

6.2. Defense in depth

Table 1 and Figure 4 compare four designs: the agent acting alone, a confidence gate, a risk-aware gate that also escalates high-risk cases the agent is not sure about, and a design that adds an independent verifier flagging high-risk proposals for human review. The unsafe-action rate falls from 0.230 with the agent alone to 0.077 with the full stack, a threefold reduction, and expected harm falls monotonically, while the escalation cost rises modestly from zero to 0.13. Each layer removes a class of failure the previous one missed, which is the essence of defense in depth [19].

Table 1. Defense-in-depth configurations. Each added layer lowers the unsafe-action rate and expected harm at a modest increase in escalation.

Design	Accuracy	Unsafe rate	Expected harm	Escalation
Agent only	0.767	0.230	0.0842	0.000
Confidence gate	0.786	0.216	0.0775	0.064
Risk-aware gate	0.791	0.114	0.0735	0.088
Verifier + human	0.797	**0.077**	**0.0692**	0.133

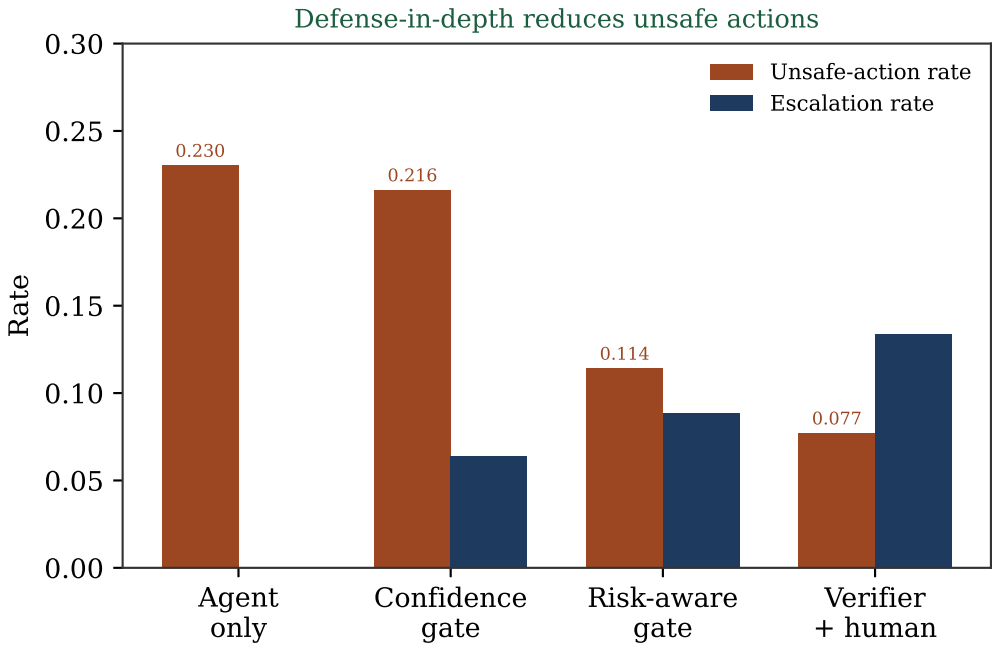

Figure 4. Defense in depth. Successive safety layers cut the unsafe-action rate from 0.230 to 0.077 while escalation rises only to 0.13.

6.3. Spending the escalation budget well

Proposition 2 predicts that a risk-weighted gate dominates a confidence-only gate at any fixed clinician budget. Figure 5 and Table 2 confirm it: across escalation budgets from two to thirty percent, ranking cases by the risk-weighted score $\rho_i(1-c_i)$ yields lower expected harm and a lower unsafe-action rate than ranking by uncertainty alone, and the gap is largest where the budget is tight and the choice of whom to escalate matters most [21].

Table 2. Risk-aware versus confidence-only escalation at matched clinician budget. The risk-weighted gate attains lower harm at every budget, as Proposition 2 predicts.

Budget	Harm (confidence)	Harm (risk-aware)	Unsafe (confidence)	Unsafe (risk-aware)
0.05	0.0792	**0.0766**	0.245	**0.181**
0.10	0.0749	**0.0710**	0.229	**0.152**
0.20	0.0667	**0.0623**	0.204	**0.128**
0.30	0.0605	**0.0542**	0.182	**0.115**

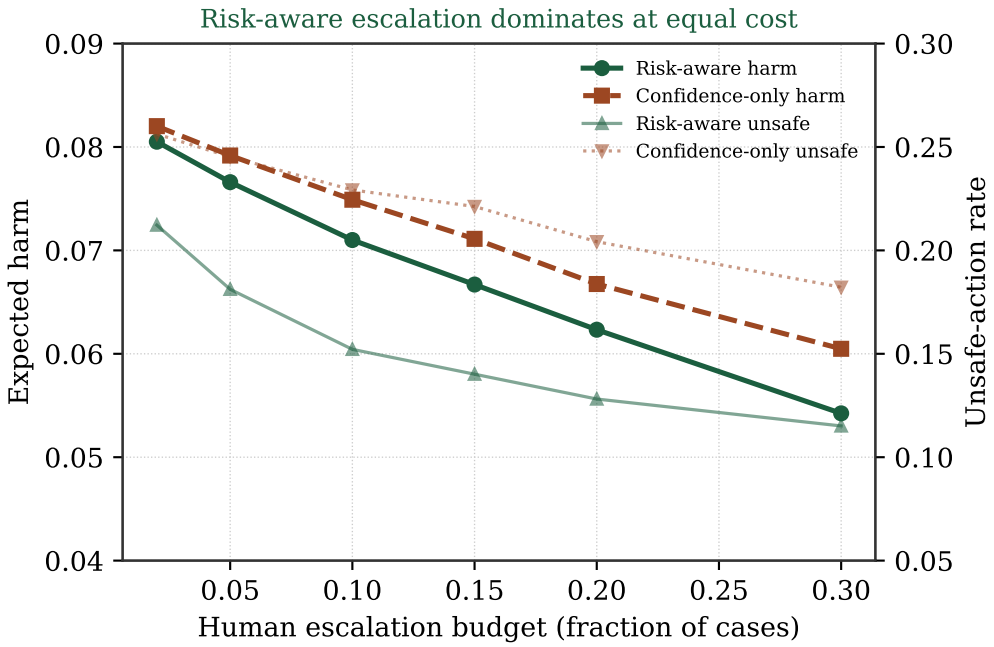

Figure 5. Spending a fixed escalation budget. A risk-weighted gate achieves lower expected harm and fewer unsafe actions than a confidence-only gate at every budget.

7. Discussion

The three studies support the design. The trade-off curve shows that autonomy and safety are exchangeable through a single tunable gate; the defense-in-depth table shows that layered checks remove unsafe actions that any one check would miss; and the budget study confirms the theory that a clinical gate should weight by risk, not by confidence alone. Together they argue that the safety of a healthcare agentic system is a property of its loop and its oversight, not only of its underlying model [10].

Several limitations bound these claims. The simulation is a model, not a clinical trial: it assumes a calibrated confidence signal and a clinician who is uniformly more accurate, and real systems violate both to a degree, which is why calibration and clinician support are themselves design targets [11]. The harm model treats risk as known; in practice ρ_i must be estimated, and a poor estimate degrades the risk-weighted gate toward the confidence-only one. Equity cannot be read off aggregate harm, since a design can lower average harm while worsening it for a subgroup, so subgroup monitoring must run alongside [21]. And explanation is not verification: an agent that can narrate a plausible rationale is not thereby safe, and the architecture leans on constraint checking and provenance rather than post hoc explanation [22]. Regulatory and interpretability considerations point the same way, toward inspectable pipelines for the highest-stakes actions [25].

Future work runs in three directions: replacing the linear agent model with a tuned clinical language model inside the same gate [9], learning the escalation policy and confidence calibration online from clinician corrections [7], and extending the audit loop into a continuous sociotechnical monitoring system that tracks drift, workload, and equity in deployment [20].

8. Conclusion

We treated healthcare agentic AI as a design problem in which a capable but fallible reasoning agent must be made safe by the system around it. We set out six design requirements from the clinical AI literature, gave a reference architecture that wraps an agentic pipeline in a verification gate and a clinician escalation path, and proved two guarantees for that gate: escalation is monotone in safety, so widening it never increases expected harm, and a risk-weighted gate is optimal for any fixed clinician budget. A fully reproducible simulation confirmed the design, showing that the gate exchanges autonomy for safety along a smooth trade-off, that layered checks cut the unsafe-action rate from 0.230 to 0.077, and that risk-weighted escalation dominates confidence-only escalation at every budget. The message for practitioners is that in medicine the safety of an agentic system is engineered at the level of the loop, through verification, risk-weighted escalation, and clinician oversight, and that these can be designed and reasoned about rather

than left to chance [15].

Acknowledgements

The authors thank colleagues and clinical collaborators for helpful discussion. All code and data needed to reproduce every figure and table are available from the authors.

References

[1] S. Singamsetty and S. Sanakkayala, "Agentic AI-driven portfolio optimization: a hybrid approach for optimized stock selection and deep learning in algorithmic trading," *Multimedia Systems*, vol. 32, no. 1, art. 70, 2026.

[2] A. Gupta, D. Nandan, S. Satyanarayana, and N. R. Rao, "Deep learning approach for brain glioblastoma detection: integrating efficient segmentation and survival rate prediction," in *Proc. 2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON)*, 2026, pp. 1–7.

[3] C. Mamatha, S. Satyanarayana, and T. K. Mohammad, "Enhanced automated system for early breast cancer prediction and classification using machine learning techniques," in *AIP Conference Proceedings*, vol. 3418, no. 1, AIP Publishing, 2026.

[4] S. Satyanarayana, T. Khatoon, and N. M. Bindu, "Breaking barriers in kidney disease detection: leveraging intelligent deep learning and artificial gorilla troops optimizer for accurate prediction," *International Journal of Applied and Natural Sciences*, vol. 1, no. 1, pp. 22–41, 2023.

[5] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: synergizing reasoning and acting in language models," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[6] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: language models can teach themselves to use tools," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 68539–68551.

[7] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: language agents with verbal reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 8634–8652.

[8] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, et al., "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, art. 186345, 2024.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[10] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[11] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah, "The shaky foundations of large language models and foundation models for electronic health records," *npj Digital Medicine*, vol. 6, art. 135, 2023.

[12] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[14] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[15] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[16] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.

[17] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, art. 299ra122, 2015.

[18] M. P. Sendak, W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, et al., "Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study," *JMIR Medical Informatics*, vol. 8, no. 7, art. e15182, 2020.

[19] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, et al., "Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality," *Journal of the American Medical Informatics Association*, vol. 10, no. 6, pp. 523–530, 2003.

[20] D. F. Sittig and H. Singh, "A new sociotechnical model for studying health information technology in complex adaptive healthcare systems," *Quality and Safety in Health Care*, vol. 19, no. Suppl 3, pp. i68–i74, 2010.

[21] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[22] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[23] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, art. 310, 2020.

[24] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 2503–2511.

[25] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.