Causal Inference-Guided Adversarial Detection and Mitigation in Medical Large Language Model Agent Architectures

Aarav Shetty
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
aarav.shetty@uc.edu

Petri L. Dawson
Department of Computer Science, George Mason University, Fairfax, VA, USA.
dawson1983@gmu.edu

Zizhan Jiang
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
zjiang@buffalo.edu

Abstract

The integration of large language models (LLMs) into autonomous agent architectures for clinical decision support, diagnostic reasoning, and patient-facing triage introduces unprecedented capabilities and simultaneous vulnerabilities. Medical LLM agents, characterized by multi-step tool-use, retrieval-augmented generation, and interactive reasoning loops, expand the attack surface for adversarial manipulation beyond single-turn prompt injection to causal perturbation of decision pathways. This paper presents a system-level framework that leverages causal inference to model, detect, and mitigate adversarial threats in medical LLM agent architectures. We argue that purely correlation-based robustness methods are insufficient for safety-critical clinical settings, as they fail to distinguish spurious associations from genuine causal mechanisms exploited by sophisticated adversaries. The proposed approach embeds causal structure learning and counterfactual reasoning within the agent's execution pipeline, enabling real-time identification of anomalous intervention patterns that deviate from clinically plausible causal graphs. We examine structural trade-offs between detection latency and clinical workflow integration, discuss infrastructure requirements for deploying causal inference modules atop existing LLM stacks, and analyze fairness implications across demographic subgroups when adversarial examples exploit historical disparities encoded in training data. The paper further addresses governance challenges, including liability attribution across agent components, continuous monitoring under distribution shift, and sustainability of computational overhead for causal inference in resource-constrained healthcare settings. Throughout, we maintain a systems perspective, emphasizing how architectural decisions about causal module integration influence robustness, interpretability, and regulatory alignment. We conclude by outlining a research agenda for causally-aware medical agent design that balances real-time performance demands with the epistemic rigor required for high-stakes clinical environments.

Keywords

causal inference, adversarial robustness, large language models, medical AI agents, system architecture, healthcare safety, counterfactual reasoning.

1. Introduction

The rapid development of large language model-based autonomous agents in medicine marks a profound shift from static predictive models to interactive systems capable of multi-step clinical reasoning, tool invocation, and dynamic retrieval of evidence. Architectures such as ReAct, Reflexion, and tool-augmented LLM pipelines now enable agents to iteratively formulate queries, consult external knowledge bases, interpret laboratory values, and generate differential diagnoses with contextual awareness [1], [2]. While these capabilities promise substantial improvements in clinical workflow efficiency and diagnostic accuracy, they simultaneously expose healthcare systems to novel adversarial threats that operate on the causal fabric of medical decision-making rather than on superficial input perturbations. Medical LLM agents, by virtue of their compositional reasoning chains and external tool dependencies, introduce attack surfaces that span prompt engineering vulnerabilities, retrieval corpus poisoning, tool output manipulation, and multi-turn interaction hijacking [3], [4].

Conventional approaches to adversarial robustness in machine learning, including adversarial training, input sanitization, and statistical outlier detection, predominantly rely on correlational patterns learned from observational data. In safety-critical clinical contexts, however, such correlational safeguards are brittle. Adversaries can craft perturbations that preserve statistical distributional properties while altering the causal structure underlying a clinical reasoning trajectory. For instance, an attack might introduce a laboratory result that, while within normal population range, causally contradicts the patient's unfolding physiological narrative in ways that a correlation-based classifier would not flag [5]. This limitation motivates the integration of causal inference frameworks into the detection and mitigation stack of medical LLM agents. Causal models explicitly represent generative mechanisms, enabling agents to reason about interventions and counterfactuals, thereby equipping them to distinguish between benign distributional shifts and malicious causal perturbations [6], [7].

This paper articulates a system-level architecture for causally-guided adversarial detection and mitigation in medical LLM agent environments. We adopt a holistic perspective, addressing not only algorithmic components but also infrastructure design, deployment trade-offs, fairness implications, governance structures, and long-term sustainability. We argue that embedding causal reasoning into the agent loop is not merely an accuracy-enhancing technical addition but a fundamental prerequisite for trustworthy autonomous clinical reasoning. The remainder of the paper is organized as follows. Section 2 reviews related work at the intersection of medical AI agents, adversarial robustness, and causal inference. Section 3 introduces a reference architecture for causally-aware medical LLM agents. Section 4 characterizes the adversarial threat landscape through a causal lens. Section 5 details detection mechanisms grounded in counterfactual consistency checks. Section 6 analyzes mitigation strategies and the structural tensions they introduce. Section 7 discusses infrastructure, deployment, and sustainability considerations. Section 8 examines fairness, governance, and policy dimensions. Section 9 concludes with forward-looking reflections.

2. Background and Related Work

Medical artificial intelligence has historically evolved through distinct eras, from rule-based expert systems to deep learning-based image classifiers, yet the emergence of LLM-powered agents constitutes the first paradigm that autonomously orchestrates complex clinical workflows. These agents decompose high-level goals into executable sub-tasks, leverage external tools such as calculators, drug databases, and guideline repositories, and maintain

stateful memory across interactions [1], [8]. In parallel, the adversarial machine learning literature has catalogued an expanding repertoire of attacks against LLMs, including jailbreaking via optimized suffixes, indirect prompt injection through retrieved documents, and gradient-based adversarial examples targeting latent representations [3], [9]. However, existing defenses primarily operate at the input-output level and lack architectural mechanisms to monitor the internal causal coherence of agent reasoning trajectories.

Causal inference, grounded in structural causal models and the do-calculus formalized by Pearl, provides a principled language for modeling interventions, confounders, and counterfactual scenarios [6]. Recent work has explored causal representation learning, causal discovery from electronic health records, and counterfactual fairness in clinical prediction [7], [10], [11]. The intersection of causality and adversarial robustness remains nascent but promising. By modeling the causal graph of a clinical scenario, an agent can assess whether a sequence of observations and actions is consistent with plausible interventional pathways. A notable contribution in this direction is the adaptation of adversarial training techniques to enforce invariance across known causal factors, yet such approaches have not been systematically integrated into the multi-modal, tool-augmented agent architectures prevalent in modern medical AI deployments [12], [13].

Furthermore, the deployment of LLM agents in medicine raises non-trivial governance questions. Regulatory frameworks from the U.S. Food and Drug Administration and the European Medicines Agency emphasize the need for continuous learning systems to incorporate risk management plans that address adversarial drift [14]. The International Medical Device Regulators Forum has highlighted the importance of causal understanding in software as a medical device to preempt catastrophic failures [15]. Our work extends these conversations by proposing a concrete systems blueprint for embedding causal inference modules into medical LLM agent stacks, with attention to the full lifecycle from design to post-market surveillance.

3. System Architecture and Causal Inference Integration

We propose a modular reference architecture for medical LLM agents augmented with causal inference capabilities. At the core, the agent comprises an LLM reasoning engine, a tool orchestration layer, a memory module, and a safety supervision unit. The causal inference module operates as a cross-cutting component that interfaces with each architectural layer. It maintains a dynamic causal graph representing the clinical case, updated as new evidence enters the system through tool outputs or user inputs. This graph is not a static Bayesian network but a continually revised structure inferred through a combination of domain knowledge, causal discovery algorithms operating on the agent's execution trace, and pre-specified clinical ontologies [6], [16].

The integration occurs at three critical junctures. First, during the observation ingestion phase, incoming data from laboratory systems, imaging reports, or patient messages are assessed for causal compatibility with the existing graph. Observations that violate conditional independence relations consistent with known physiological mechanisms trigger elevated scrutiny. Second, the tool selection and planning stage uses causal reasoning to validate that a proposed action—such as ordering a specific diagnostic test—is causally justified given the patient's state and clinical guidelines. Third, the output generation stage applies counterfactual checks: the agent generates implicit counterfactuals ("what would this diagnosis imply if symptom X were absent?") and evaluates coherence with the current

evidence graph. This multi-phase integration ensures that adversarial perturbations are caught before they can propagate through the agent's decision loop.

The architecture introduces several structural trade-offs. Embedding causal reasoning increases computational complexity, with graph update operations requiring non-trivial probabilistic inference over potentially high-dimensional clinical variables. This latency must be balanced against the real-time demands of emergency department triage or intraoperative decision support. One design choice involves decoupling the causal module into an asynchronous verifier that operates on a slight temporal delay, flagging inconsistencies for human review rather than blocking agent actions outright. Such a configuration maintains low latency for critical pathways while still providing a causal safety net. Another trade-off concerns the granularity of the causal graph: fine-grained representations capturing detailed physiological mechanisms offer higher detection accuracy but demand substantial domain engineering and computational resources, whereas coarser graphs derived from clinical guidelines are more scalable but potentially miss subtle adversarial manipulations that exploit edge-case causal interactions [17].

4. Adversarial Threat Landscape in Medical LLM Agents

Adversaries targeting medical LLM agents can exploit the multi-step, tool-augmented nature of these systems in ways that differ fundamentally from single-turn attacks on language models. We categorize threats along two axes: the stage of the agent loop targeted, and the causal depth of the perturbation. Input-stage attacks include modified patient histories, injected contradictory findings in retrieval corpora, and adversarially crafted radiology report summaries. Mid-loop attacks manipulate the outputs of tools called by the agent, such as laboratory value APIs or drug interaction databases, after the initial reasoning step has committed to a causal direction. Output-stage attacks aim to bias the final recommendation without altering intermediate observations, for example by embedding subtle logical fallacies that the summary generator amplifies.

The clinical impact of even statistically subtle perturbations can be severe when they disrupt causal chains. Consider an agent managing a patient with suspected sepsis. An adversary that lowers the reported white blood cell count by a small, clinically plausible margin might not trigger any anomaly detector based on population norms. However, in the causal context of the patient's fever, tachycardia, and recent surgery, the manipulation invalidates the expected causal pathway linking infection to leukocytosis, potentially delaying antibiotic administration. A causally-aware detector, by contrast, would note that the joint distribution of symptoms and laboratory values under the sepsis causal model assigns extremely low likelihood to that particular pattern, regardless of how common the individual values are in the general population.

Tool-output corruption poses particularly insidious risks because tools are often treated as oracles within agent architectures [18]. An attacker who compromises the drug information database could introduce false contraindications that redirect the agent toward less effective therapies, all while adhering to surface-level safety constraints. Multi-agent medical systems, where separate agents handle triage, diagnosis, and treatment planning, introduce additional cascading failure modes where a single corrupted causal link propagates across handoff boundaries. The temporal dimension further complicates detection, as adversaries can drip-feed subtle causal inconsistencies over multiple interactions, gradually steering the agent's latent state toward a dangerous basin of attraction without any single step appearing overtly malicious.

5. Causal Inference-Guided Detection Mechanisms

Our framework's detection strategy relies on monitoring causal consistency throughout the agent's execution trace. The core mechanism is a causal discrepancy score calculated as the divergence between the observed joint distribution of the execution trace variables and the distribution implied by the current causal graph under the assumption of no interventions. At each agent step, the causal module updates its internal graph representation using both prior clinical knowledge and the incoming data's conditional independence structure. It then performs a suite of counterfactual consistency checks: for key clinical variables, it computes the expected outcomes under independently perturbed inputs and tests whether the actual agent behavior aligns with the causal model's predictions [6], [19].

For example, in a diagnostic reasoning task, the detector might ask: "Given that the agent proposed diagnosis D and observed symptom S1, what is the expected probability of S1 had S2 been normal?" If the agent's internal attention patterns or probability outputs contradict the causal independence relations encoded in the medical domain graph, a potential adversarial perturbation is flagged. Importantly, these checks are implemented without requiring access to ground-truth diagnoses, making them deployable in real-world unsupervised monitoring roles. To reduce false positive rates, the system combines causal consistency scores with epistemic uncertainty estimates from the LLM, recognizing that high uncertainty regions may naturally exhibit causal anomalies even in benign settings [20].

A complementary detection method uses causal mediation analysis on the agent's action sequences. By estimating the natural direct and indirect effects of upstream observations on final recommendations, the system identifies whether any tool output or retrieved document exerts an outsized causal influence inconsistent with its clinical role. This approach has proven useful in financial fraud detection and can be adapted to the clinical domain by leveraging established effect size norms from evidence-based medicine [21]. Detection thresholds are calibrated on curated benign and adversarial clinical vignettes, with ongoing adaptation to local institutional practice patterns to prevent distributional drift from undermining sensitivity.

6. Mitigation Strategies and Structural Trade-offs

Upon detecting a causal anomaly, the agent invokes mitigation protocols that range from low-certainty alerts to hard action blocks. The choice of response introduces critical architectural trade-offs. Immediate action suppression ensures safety but may disrupt time-sensitive clinical workflows, whereas passive flagging for later review preserves throughput at the risk of allowing harm if the attack operates on a shorter timescale than human oversight. We advocate a graduated response framework conditioned on both the causal discrepancy magnitude and the clinical criticality of the decision. High-stakes decisions, such as initiating insulin therapy or administering thrombolytics, trigger obligatory human-in-the-loop confirmation irrespective of causal confidence, aligning with the precepts of meaningful human control in high-risk AI systems [22].

A structural challenge emerges in balancing adversarial robustness against natural distributional shift. Medical practice evolves as new evidence emerges and guidelines are revised. A causal model that is too rigidly anchored to a static domain graph will generate false positives when legitimate clinical innovations introduce novel causal pathways. Conversely, a model that adapts too readily to observed patterns risks absorbing adversarial manipulations into its causal structure, treating them as authentic clinical discoveries. We

propose a semi-supervised causal graph refinement process where updates to the domain graph require validation by a human clinical board at scheduled intervals, while low-stakes edges can adapt continuously under statistical constraints that limit the influence of any single user interaction [23].

The interplay between causal mitigation and LLM fine-tuning also merits attention. If an agent is fine-tuned on inclusive clinical data to reduce demographic bias, the causal graph must simultaneously be updated to reflect population-specific causal directions, such as varying symptom presentations across ethnic groups. Adversaries aware of fine-tuning schedules could time attacks to coincide with model updates, exploiting transitional periods where the causal module and the policy model are misaligned. Versioned causal graph registries and staged rollout protocols, akin to those used in safety-critical software updates for medical devices, can mitigate this synchronization vulnerability.

7. Infrastructure, Deployment, and Sustainability

Deploying causal inference-guided medical LLM agents in real healthcare environments demands infrastructure capable of supporting low-latency graph inference alongside compute-intensive transformer forward passes. While dedicated hardware acceleration and optimized causal inference libraries have reduced computational costs, the aggregate overhead remains significant for hospitals with constrained IT budgets. A potential solution involves tiered deployment architectures where a centralized causal model serving layer, hosted in a hospital's private cloud or edge facility, handles graph updates and counterfactual queries for multiple agent instances. This centralization also facilitates consistent causal graph versioning and audit logging for regulatory compliance.

Sustainability considerations extend beyond energy consumption to the maintenance burden of curated clinical causal graphs. Constructing comprehensive causal models for even a single medical specialty requires sustained collaboration between clinicians, ontologists, and machine learning engineers. To mitigate this, we suggest leveraging pre-existing biomedical knowledge graphs and the outputs of causal structure learning algorithms as bootstrap components, with incremental manual refinement focused on high-risk clinical scenarios [24]. The long-term viability of such systems depends on establishing shared community resources similar to the Unified Medical Language System but extended with causal annotations, maintained by professional societies and public health agencies.

Interoperability with electronic health record systems introduces additional complexity. Real-time causal checks require access to structured clinical data, which remains fragmented across proprietary vendor schemas despite ongoing standardization efforts. Adopting Fast Healthcare Interoperability Resources with causal metadata extensions could enable plug-and-play integration of causal monitoring modules across different institutional environments. Furthermore, deployment in telemedicine settings, where LLM agents increasingly serve as initial patient-facing triage, introduces network latency and bandwidth constraints that necessitate lightweight causal models, perhaps using amortized inference techniques to reduce runtime cost while preserving causal rigor.

8. Fairness, Governance, and Policy Implications

Adversarial attacks on medical LLM agents can asymmetrically affect marginalized patient populations by exploiting existing disparities in clinical data. If attackers target causal pathways that are already poorly calibrated for underrepresented groups—such as modifying symptom prompts that leverage race-specific normal ranges—the resulting harms compound

preexisting inequities. Causal fairness frameworks, which enforce equalized counterfactual outcomes across protected attributes, offer a principled defense but require careful specification of admissible and inadmissible causal pathways [10], [25]. A causally-aware detection system can be designed to raise sensitivity for causal pattern disruptions that correlate with demographic attributes, though this raises the spectre of differential false positive rates that could erode trust among vulnerable communities.

Governance structures must evolve to address the distributed nature of medical LLM agent supply chains. When an adversarial attack succeeds, liability may be contested among the LLM provider, the tool integration vendor, the healthcare institution, and the clinical supervisor. We posit that causal audit trails, stored immutably and indexing every graph update with its provenance, can support forensic investigations and clarify accountability. Regulatory frameworks such as the European Union's Artificial Intelligence Act's requirements for high-risk AI systems to incorporate explainability and robustness can be satisfied, in part, by demonstrating that causal monitoring was operational and triggered defined mitigation protocols [15]. However, standards bodies must develop specific causal auditing criteria, as current guidelines remain at an abstract level that offers limited conformance testing guidance.

International harmonization of medical AI governance remains fragmented, with jurisdictions adopting divergent risk classification and post-market surveillance requirements. A global framework for causal robustness certification, perhaps modeled on aviation safety's shared incident reporting systems, could accelerate learning across health systems while respecting data sovereignty constraints. Policy incentives, including reimbursement models that reward proactive safety investments rather than volume-based payments, would accelerate adoption of computationally expensive causal inference infrastructures, aligning economic drivers with patient safety imperatives.

9. Conclusion

This paper has argued that causal inference provides an indispensable lens for detecting and mitigating adversarial threats in medical LLM agent architectures. By embedding causal consistency checking and counterfactual reasoning into the agent execution loop, systems can transcend the limitations of correlation-based robustness methods and address the unique vulnerabilities introduced by multi-step, tool-augmented clinical reasoning. We have presented a reference architecture that integrates causal modules across observation, planning, and generation phases, discussed structural trade-offs between latency, safety, and scalability, and explored the infrastructure and governance prerequisites for real-world deployment. The fairness analysis underscored the risk that adversarial exploitation of underrepresented causal pathways could deepen health inequities, while the policy discussion highlighted the need for causal audit trails and harmonized international standards. Moving forward, the medical AI research community must invest in open causal graph resources, standardized adversarial evaluation benchmarks for clinical agent pipelines, and longitudinal field studies that measure the real-world efficacy of causally-guided detection in reducing preventable adverse events. The convergence of causal inference and LLM agent architectures holds the promise of elevating medical AI from a pattern-recognition tool to a trustworthy clinical reasoning partner, but only if systems-level design principles are embraced with the same rigor that has guided safety engineering in other high-consequence industries.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.

3. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

4. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv preprint arXiv:2302.12173.

5. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

6. Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

7. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612–634.

8. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, Y., ... & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

9. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (pp. 39–57). IEEE.

10. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30.

11. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

12. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

13. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.

14. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.

15. International Medical Device Regulators Forum. (2022). Software as a medical device: Possible framework for risk categorization and corresponding considerations. IMDRF/SaMD WG/N12.

16. Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. MIT Press.

17. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: Foundations and learning algorithms. MIT Press.

18. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

19. Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

20. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning (pp. 1050–1059). PMLR.

21. Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

22. Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of "autonomous systems". IEEE Intelligent Systems, 28(3), 54–59.

23. Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27), 7345–7352.

24. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(1), D267–D270.

25. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.