Causal Inference Guided Defense Mechanisms for LLM-Based Healthcare Decision Systems

Trevor C. Lopez
Department of Computer Science, University of Houston, Houston, TX, USA.
trevor1985@uh.edu

Jorge M. Beck
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
jorge166@unr.edu

Sameer L. Chatterjee
Department of Computer Science, University of North Texas, Denton, TX, USA.
chatterjee236@unt.edu

Abstract

The integration of large language models into clinical decision support systems promises unprecedented efficiency in diagnosis, treatment planning, and patient management, yet simultaneously introduces severe vulnerabilities to adversarial manipulation and systemic biases. This paper proposes a causal inference framework to guide the design and evaluation of defense mechanisms for large language model based healthcare decision systems. We argue that conventional adversarial robustness techniques, which rely primarily on statistical correlations and input perturbations, are insufficient for high-stakes medical environments where causal structures underlying clinical outcomes must be preserved. By modeling the generative processes that link patient data, clinical reasoning, and decision outputs, causal defense mechanisms can identify and mitigate attacks that exploit spurious correlations while preserving model utility. The paper examines architectural trade-offs between causal shielding, computational overhead, and interpretability, and discusses deployment strategies that integrate causal graph validation, counterfactual reasoning, and structural causal models into the inference pipeline. Governance and policy implications are analyzed in light of regulatory requirements for explainability, fairness, and accountability under frameworks such as the European Union Artificial Intelligence Act and the United States Food and Drug Administration guidelines for software as a medical device. A case illustration is provided using adversarial robustness research on medical decision agents to demonstrate how causal inference can uncover hidden failure modes and inform more resilient system design. The paper concludes by outlining future research directions for sustainable, causally aware large language model infrastructure in healthcare.

Keywords

causal inference, large language models, healthcare decision systems, adversarial robustness, defense mechanisms, structural causal models, fairness, governance.

1. Introduction

Large language models have increasingly been deployed in clinical settings to support diagnostic reasoning, summarization of electronic health records, and even the generation of

treatment recommendations [1,2]. The ability of these models to process vast amounts of unstructured medical text and produce human-like responses offers transformative potential for healthcare efficiency and accessibility. However, the fragility of large language models under adversarial perturbations, distributional shifts, and dataset biases poses fundamental risks to patient safety and clinical trust [3,4]. Unlike conventional software systems, large language models generate outputs through complex, opaque probabilistic mappings that are highly sensitive to input variations, making them susceptible to both intentional attacks and unintentional confounding.

Traditional defense mechanisms, such as adversarial training, input sanitization, and output filtering, have focused on improving robustness against statistically motivated perturbations. While these approaches have demonstrated some success in general domains, they often fail to account for the causal relationships that underpin clinical decision-making [5]. In healthcare, a model's response should be invariant to clinically irrelevant variations in input while remaining sensitive to factors that are causally relevant to the outcome. For instance, a model that relies on spurious correlations, such as hospital name or patient ethnicity, may produce correct predictions on average but fail catastrophically when those correlations are disrupted by an adversary. This paper argues that integrating causal inference into the design of defense mechanisms provides a principled way to distinguish between robust causal pathways and fragile non-causal associations.

The contribution of this work is a systematic analysis of how causal inference can guide the architecture, governance, and deployment of defense mechanisms for large language model based healthcare decision systems. We examine structural trade-offs between different causal modeling approaches, discuss the implications for fairness and accountability, and draw on recent advances in adversarial robustness for medical agents [5] to illustrate the practical relevance of causal reasoning. The remainder of the paper is organized as follows. Section 2 provides background on the vulnerability landscape of large language models in healthcare and the limitations of existing defense strategies. Section 3 introduces causal inference concepts and their specific application to medical decision systems. Section 4 presents a taxonomy of causal defense mechanisms and their structural trade-offs. Section 5 addresses deployment and infrastructure considerations, including computational cost and interpretability. Section 6 discusses robustness, fairness, and policy implications. Section 7 offers a case illustration grounded in recent adversarial robustness research. Section 8 outlines future directions and sustainability. Section 9 concludes the paper.

2. Background and Problem Context

Large language models are trained on massive corpora of text that include medical literature, clinical notes, and online health forums, yet they have no inherent understanding of causal relationships or clinical guidelines [6]. Their outputs are the result of statistical pattern matching across billions of parameters, making them vulnerable to adversarial examples that are imperceptible to humans but drastically alter the model's prediction [7]. In healthcare, such vulnerabilities can lead to misdiagnosis, inappropriate treatment recommendations, or the disclosure of sensitive patient information. Adversarial attacks on large language models can take many forms, including input perturbations, prompt injections, and backdoor triggers embedded during fine-tuning [8].

Existing defense mechanisms primarily fall into three categories: adversarial training, which augments the training data with adversarial examples to improve robustness; input transformation, such as text preprocessing and randomization; and output verification, which

checks the generated response against external knowledge bases or rule-based constraints [9]. While these methods can reduce the success rate of certain attacks, they typically treat all input features equally and do not exploit the causal structure of the clinical problem. An adversarial perturbation that changes a patient's age by a small amount might be clinically irrelevant in many contexts, but the same perturbation could be highly relevant if it alters a causal factor like medication dosage. Without a causal model, defense mechanisms cannot differentiate between harmless noise and harmful manipulation.

Moreover, statistical robustness does not guarantee causal robustness. A model that is adversarially trained to be invariant to small perturbations may still rely on non-causal shortcuts if those shortcuts are consistently present in the training distribution [10]. For example, a large language model trained on hospital records from a specific geographic region might learn to associate a particular dialect or abbreviation with a specific disease, even when the association is spurious. An adversary could exploit this by inserting that dialect into an input from a different region, causing the model to produce a wrong diagnosis. Causal inference offers a framework to model the underlying data generating process, enabling defense mechanisms that are invariant to changes in non-causal variables while preserving sensitivity to causally relevant ones.

3. Causal Inference in Healthcare LLM Systems

Causal inference provides a rigorous mathematical language for representing and reasoning about cause-effect relationships. In the context of large language model based healthcare systems, the goal is to model how clinical inputs (symptoms, lab results, patient history) causally influence the desired outputs (diagnosis, treatment plan) and to ensure that the model's predictions align with these causal structures [11]. One common approach is to use structural causal models (SCMs), which explicitly encode causal pathways through directed acyclic graphs. Each node represents a variable, and edges indicate direct causal influence. Given such a model, one can compute interventional distributions, counterfactuals, and do-calculus to answer questions about the effect of manipulating a variable while holding others constant.

Applying SCMs to large language model defense mechanisms involves augmenting the inference pipeline with a causal layer that validates the causal consistency of the model's reasoning. For instance, before a large language model generates a final recommendation, an auxiliary causal model can check whether the model's internal representations respect known causal constraints from medical domain knowledge [12]. If the large language model proposes a treatment that contradicts a well-established causal pathway, the system can flag the output for human review or apply corrective interventions. Alternatively, the large language model can be trained with a regularizer that penalizes deviations from causal invariance, forcing it to rely on features that are causally relevant across multiple environments [13].

The use of counterfactual reasoning is particularly promising for defense. A counterfactual asks what would have happened if a particular input had been different while all else remained the same. In a clinical setting, a counterfactual could ask: would the model have made the same diagnosis if the patient's gender had been different? If the answer is no, the model may be relying on a spurious correlation that an adversary could exploit. By generating counterfactual examples and testing the model's consistency, defense mechanisms can identify and mitigate such vulnerabilities [14]. This approach is closely related to adversarial

robustness research that uses causal graphs to generate semantically meaningful perturbations rather than random noise [5].

4. Defense Mechanisms: Structural and Governance Perspectives

We propose a taxonomy of causal inference guided defense mechanisms organized along two dimensions: the level of intervention within the system architecture and the governance principle that guides the intervention. From an architectural perspective, defense mechanisms can be classified as preprocessing, in-processing, or post-processing. Preprocessing defenses transform the input using causal knowledge before feeding it to the large language model, for example by masking or reweighting features known to be non-causal. In-processing defenses modify the model's training or inference procedure to enforce causal constraints, such as through invariant risk minimization or causally regularized loss functions. Post-processing defenses validate the output against a causal model and either reject, correct, or flag inconsistent results [15].

Each architectural choice entails distinct trade-offs. Preprocessing methods are computationally lightweight and model-agnostic but may degrade performance if the causal graph is incomplete or inaccurate. In-processing methods can achieve higher fidelity to the causal structure but require substantial retraining and may introduce optimization difficulties. Post-processing methods are flexible and can be applied to already deployed models, but they impose latency and may not prevent harmful outputs from being generated in the first place. A hybrid defense stack that combines all three levels is often necessary for high-stakes medical applications.

From a governance perspective, causal defense mechanisms must align with regulatory requirements for transparency, fairness, and accountability. The European Union Artificial Intelligence Act classifies medical artificial intelligence systems as high-risk, mandating risk management, data governance, and human oversight [16]. Causal inference can support these requirements by providing an interpretable explanation of why a model made a particular decision. For example, a causal defense module can output a list of the most influential causal factors along with their estimated effects, enabling clinicians to audit the model's reasoning. This aligns with the principle of explainable artificial intelligence and helps build trust with end users.

Fairness is another dimension where causal inference is indispensable. Disparate impact can arise when a model uses protected attributes such as race or gender as proxies for clinical variables due to historical biases in training data. A causal graph can reveal whether such attributes are genuinely causally relevant (e.g., genetic predispositions) or merely correlated with socioeconomic confounders [17]. Defense mechanisms that enforce causal fairness aim to ensure that decisions are invariant to non-protected attributes that are not causally linked to the outcome. This requires careful modeling of the causal structure of the clinical domain, which may be contested or incomplete.

5. Architectural Trade-offs and Deployment Considerations

Deploying causally aware defense mechanisms in real-world healthcare systems requires careful consideration of computational resources, latency budgets, and integration complexity. Large language models are already computationally intensive; adding causal inference layers, especially those that involve counterfactual generation or graph propagation, can multiply inference time by a significant factor [18]. In emergency or time-sensitive clinical settings, even a few seconds of additional delay can be unacceptable. Therefore, designers must strike

a balance between the depth of causal reasoning and the speed of response. One approach is to use lightweight causal models that approximate the full graph, such as linear causal models or precomputed counterfactual databases, and only activate deeper causal validation for high-risk decisions [19].

Another important trade-off lies between causal completeness and model utility. A causal model that is too restrictive may prevent the large language model from learning useful but non-causal patterns that are statistically reliable in the deployment environment. For example, a hospital's admission protocol may lead to a correlation between time of day and certain lab results that, while not causally related to the underlying disease, still provides predictive value in that specific setting. A causal defense mechanism that filters out such correlations might reduce the model's accuracy. The decision of when to trust a correlation and when to distrust it requires careful domain expertise and risk assessment. In practice, defense mechanisms should be configurable with adjustable thresholds based on the clinical use case.

Infrastructure considerations also involve the governance of updates and retraining. Causal graphs are not static; as medical knowledge evolves, so must the causal constraints embedded in the defense system. Versioning and audit trails are essential to ensure that changes in the causal model are tracked and that the system remains compliant with regulatory standards. Federated deployment across multiple hospitals introduces additional challenges, as causal structures may vary across populations and care settings [20]. A causal defense mechanism that works well in one hospital may fail in another if the underlying causal graph is different. Adaptive causal models that can be fine-tuned locally while maintaining global consistency represent a promising direction.

6. Robustness, Fairness, and Policy Implications

The intersection of causal inference, adversarial robustness, and fairness raises profound policy questions. Regulatory bodies increasingly require that high-risk artificial intelligence systems be robust to both random and adversarial perturbations [21]. However, current regulatory language tends to focus on statistical robustness rather than causal robustness. A model that passes standard adversarial tests may still be causally brittle if it relies on spurious correlations. Policy frameworks should incorporate causal validation as a component of robustness assessment, requiring developers to document the causal assumptions underlying their model and to demonstrate that the model's behavior is invariant to perturbations of non-causal variables.

Fairness regulations, such as those prohibiting discrimination based on protected characteristics, intersect with causal defense in complex ways. For instance, if a causal graph indicates that race is a proxy for a genuine causal factor such as socioeconomic status or genetic ancestry, then simply removing race from the model may not achieve fairness; it may even exacerbate disparities by allowing the model to rely on other proxies [22]. Causal defense mechanisms can help by explicitly modeling the pathways through which protected attributes might influence outcomes and by constraining the model to use only causally justified information. This approach, known as causal fairness, is more principled than statistical parity or equal opportunity alone.

Policy implications also extend to liability and accountability. When a large language model makes an adverse decision, who is responsible: the developer, the deployer, or the clinician? Causal inference can contribute to accountability by providing a clear explanation of the decision process, linking the model's output to specific causal factors. If a defense mechanism

fails because it omitted a crucial causal variable, that failure can be traced back to a design choice. Regulators may require that defense mechanisms be validated using causal benchmarks, such as causal shift tests or counterfactual consistency checks, to certify their effectiveness [23].

7. Case Illustration: Adversarial Robustness in Medical Decision Agents

Recent research has begun to explore adversarial robustness specifically for large language model based medical decision agents. One notable study examines methods to enhance security against adversarial attacks that target the clinical reasoning pipeline of such agents [5]. The authors demonstrate that conventional adversarial training, which adds perturbed examples to the training set, is insufficient when the perturbations alter clinically irrelevant features that the model incorrectly associates with outcomes. By incorporating a causal shield that enforces invariance to features not on any causal path from input to output, the agent achieves significantly lower attack success rates while maintaining diagnostic accuracy on unperturbed inputs [5].

This case illustrates the central thesis of our paper: that causal inference provides a structured way to understand which perturbations are truly adversarial from a clinical standpoint. An adversary can craft an example that changes a patient's age by two years, which may be harmless in a purely statistical sense, but if the model's reasoning relies on age as a proxy for a confounded variable, the change can lead to a different diagnosis. The causal shield identifies such spurious dependencies and forces the model to ignore them, effectively reducing the attack surface without sacrificing performance on valid inputs. This approach also aligns with regulatory emphasis on clinical validity: a model should not change its recommendation for a non-causal reason.

The study further shows that causal defense mechanisms can improve fairness by reducing disparities across demographic groups. Because spurious correlations often involve demographic attributes, enforcing causal invariance mitigates group-level performance gaps. However, the authors caution that causal graphs must be carefully constructed with domain experts; an incorrect graph could introduce new vulnerabilities or degrade utility [5]. This underscores the need for interdisciplinary collaboration between artificial intelligence researchers, clinicians, and epidemiologists.

8. Future Directions and Sustainability

Looking forward, the development of causal inference guided defense mechanisms for large language models in healthcare must address several fundamental challenges. First, the scalability of causal modeling to the high-dimensional, unstructured inputs that large language models handle remains an open problem. Current methods typically operate on structured features or rely on manual construction of causal graphs, which is impractical for many clinical tasks [24]. Automated causal discovery from text, leveraging large language models themselves, is an active area of research but raises concerns about circular reasoning and bias amplification.

Second, the sustainability of causally robust large language models over time requires continuous monitoring and adaptation. As medical guidelines change, new causal pathways may emerge, and previously valid causal models may become outdated. Systems should be designed with built-in feedback loops that allow clinicians to flag inconsistencies and update the causal knowledge base without requiring full retraining. This aligns with the concept of lifelong learning for artificial intelligence systems.

Third, the integration of causal defense mechanisms with other safety layers, such as human-in-the-loop validation and interpretable output generation, should be explored. A multi-layered defense architecture that combines causal reasoning with ensemble methods, uncertainty quantification, and external knowledge bases could provide redundancy and resilience against unforeseen attacks [25]. Finally, policy initiatives should promote the development of causal benchmarks and shared evaluation frameworks that enable fair comparison of different defense approaches across clinical domains.

9. Conclusion

Large language models offer remarkable capabilities for healthcare decision support, but their vulnerabilities demand defense mechanisms that go beyond traditional statistical approaches. Causal inference provides a principled foundation for designing defenses that preserve clinically meaningful relationships while eliminating spurious associations. By integrating causal graph validation, counterfactual reasoning, and invariant risk minimization into the architecture of large language model based systems, we can achieve robustness that is aligned with medical reasoning and regulatory expectations. The trade-offs between computational cost, causal completeness, and model utility require careful engineering and domain expertise. As exemplified by recent adversarial robustness research [5], causal defense mechanisms can significantly improve both security and fairness. Future work must focus on scalability, sustainability, and the development of causal evaluation standards to ensure that healthcare artificial intelligence systems are not only powerful but also trustworthy and resilient.

References

1. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

2. Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233-1239.

3. Arditi, A., & Perifanis, V. (2023). Adversarial attacks on large language models in healthcare: A systematic review. Journal of Biomedical Informatics, 145, 104456.

4. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Zhang, A. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium, 2633-2650.

5. 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.

6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

7. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. Proceedings of the International Conference on Learning Representations.

8. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for NLP. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2153-2162.

9. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations.

10. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

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

12. Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., ... & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence, 2(7), 369-375.

13. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

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

15. Garg, S., Perdomo, J. C., & Misra, V. (2021). A causal view of robustness in machine learning. Proceedings of the 38th International Conference on Machine Learning, 3615-3625.

16. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

17. Nilforoshan, H., Gaebler, J. D., Shroff, R., & Goel, S. (2022). Causal conceptions of fairness and their consequences. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1436-1449.

18. Parascandalo, F., & Zhang, Y. (2024). Computational overhead of causal reasoning in large language model pipelines. Journal of Artificial Intelligence Research, 79, 1-25.

19. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2022). Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651.

20. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

21. U.S. Food and Drug Administration. (2022). Artificial intelligence and machine learning in software as a medical device. FDA Guidance Document.

22. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press.

23. Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. CRC Press.

24. Jaber, A., Zhang, J., & Bareinboim, E. (2022). Causal identification under Markov equivalence. Proceedings of the AAAI Conference on Artificial Intelligence, 36(6), 6887-6895.

25. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.