Cross-Modal Adversarial Defense for Medical LLM Agents in Clinical Decision Support Systems

Leif Willis
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
leifmail@missouri.edu

Abstract

The integration of large language model (LLM) agents into clinical decision support systems (CDSS) represents a transformative advance in healthcare informatics, yet it introduces unprecedented vulnerabilities to cross-modal adversarial attacks. Such attacks exploit the interplay between textual, visual, and structured clinical data to degrade model performance, mislead diagnostic reasoning, and potentially harm patient outcomes. This paper develops a comprehensive framework for cross-modal adversarial defense tailored to medical LLM agents operating within critical care environments. We examine the architectural foundations of multimodal LLM agents, the unique threat surfaces arising from heterogeneous input channels, and the structural trade-offs inherent in designing robust defense mechanisms. A system-level perspective is adopted to evaluate governance constraints, deployment infrastructure, sustainability of defenses under distributional shift, and fairness implications for diverse patient populations. Drawing from adversarial machine learning theory, clinical safety standards, and socio-technical infrastructure design, we propose a multi-layered defense strategy that integrates input sanitization, cross-modal consistency verification, robust training procedures, and real-time anomaly detection. The analysis highlights that no single defense suffices; rather, a coordinated ecosystem of technical safeguards, policy frameworks, and institutional oversight is necessary. We further discuss the scalability of these defenses across different healthcare settings, from tertiary hospitals to resource-constrained clinics, and consider the ethical and regulatory dimensions of deploying adversarial defenses that may inadvertently introduce biases. This work contributes to the emerging field of safe and trustworthy medical AI by providing a systematic roadmap for defending multimodal LLM agents in high-stakes clinical decision support.

Keywords

adversarial defense, cross-modal, large language models, clinical decision support, medical AI, robustness, fairness, governance.

1. Introduction

Clinical decision support systems have long sought to augment physician judgment with algorithmic insights, but the recent emergence of large language models capable of processing text, images, and structured data simultaneously has opened new possibilities for comprehensive patient assessment. Medical LLM agents can interpret radiology reports alongside imaging data, integrate laboratory results with narrative clinical notes, and generate differential diagnoses that synthesize disparate information sources. However, the very multimodality that enables these capabilities also expands the attack surface for adversarial manipulation. An adversary may inject imperceptible perturbations into a medical image, subtly alter the wording of a clinical note, or manipulate structured laboratory fields to cause

the LLM agent to reach incorrect or dangerous conclusions. Cross-modal adversarial attacks are particularly insidious because the mismatch between modalities can be exploited to bypass defenses designed for unimodal inputs. For instance, a perturbation in the image may be undetectable to a text-only filter, while a textual corruption may evade image-based detection.

The need for robust cross-modal defense is not merely an academic concern. Real-world deployment of medical LLM agents in hospitals, telemedicine platforms, and clinical research settings demands that these systems withstand deliberate attempts to undermine their reliability. Malicious actors could include disgruntled employees, external hackers seeking to cause harm or to manipulate insurance claims, or even well-intentioned users inadvertently providing corrupted data. Moreover, the stochastic nature of LLMs means that even minor perturbations can cascade into significant errors in reasoning. Therefore, a systematic understanding of adversarial vulnerabilities and corresponding defense architectures is essential for responsible integration of AI into clinical workflows.

This paper provides a comprehensive examination of cross-modal adversarial defense for medical LLM agents, adopting a system-level perspective that spans technical architecture, governance, deployment infrastructure, and policy implications. We begin by reviewing the foundational concepts of adversarial attacks in multimodal settings and the specific characteristics of medical LLM agents. Subsequently, we analyze the structural trade-offs inherent in different defense strategies, emphasizing how choices about computational cost, latency, interpretability, and fairness interact with clinical requirements. We then discuss governance frameworks that can ensure defenses remain effective over time, including continuous monitoring, update protocols, and human-in-the-loop oversight. Infrastructure considerations such as data storage, model serving, and edge deployment are examined with respect to sustainability and resilience. Finally, we address the critical issues of robustness across demographic groups and the potential for defenses to introduce or exacerbate disparities. The paper concludes with a forward-looking research agenda that prioritizes both technical innovation and socio-technical alignment.

2. Background and Related Work

Adversarial machine learning has matured significantly over the past decade, with seminal works demonstrating the vulnerability of deep neural networks to small, carefully crafted input perturbations [1]. These attacks have been extended to multimodal systems, where perturbations can be applied independently to each modality or coordinated across modalities to maximize misclassification [2]. In the medical domain, adversarial examples have been shown to fool diagnostic models for chest X-rays, histopathology slides, and skin lesions [3], raising concerns about patient safety. More recently, the rise of LLMs has introduced new attack vectors, including prompt injection, jailbreaking, and data poisoning, which can be amplified when combined with visual inputs [4]. Medical LLM agents, such as those fine-tuned on clinical corpora and integrated with vision encoders, inherit these vulnerabilities while adding domain-specific risks.

Cross-modal defenses have been explored in computer vision and NLP, often relying on adversarial training, input preprocessing, or detection mechanisms [5]. However, these approaches are typically unimodal in scope. Multimodal consistency checks, where outputs from different modalities are compared to detect anomalies, have shown promise in speech and image tasks [6], but their application to clinical data requires careful adaptation due to the inherent uncertainty and variability in medical information. For instance, a patient's

symptoms described in text may not perfectly align with imaging findings even in benign cases, making consistency thresholds difficult to set.

Governing bodies such as the U.S. Food and Drug Administration (FDA) have begun to issue guidelines for AI-based medical devices, emphasizing the need for robustness and validation across distribution shifts [7]. Yet, these frameworks do not explicitly address cross-modal adversarial attacks. The literature on fairness in medical AI highlights that adversarial defenses may disproportionately affect underrepresented groups if the adversarial perturbations or defense mechanisms correlate with demographic factors [8]. For example, a defense that relies on detecting unusual pixel patterns might flag more often for patients with darker skin tones if the training data are biased.

The specific challenges of deploying LLM agents in clinical decision support have been discussed from an infrastructural perspective, with attention to latency, interpretability, and trust [9]. However, the intersection of adversarial defense and deployment constraints remains underexplored. Our work builds on these foundations by synthesizing cross-modal attack theory, clinical safety requirements, and socio-technical governance into a unified framework.

3. Cross-Modal Vulnerabilities in Medical LLM Agents

Medical LLM agents typically consist of a large language model backbone, one or more vision encoders for medical imaging, and a module for integrating structured data such as laboratory values, vital signs, and genomic information. The fusion of these modalities can occur at the input level, intermediate representation level, or output level. Each fusion architecture introduces distinct vulnerabilities. In early fusion, where raw or minimally processed inputs are concatenated, an adversary can perturb a single modality to corrupt the joint representation. For example, an attack on a chest X-ray image could shift the embedding in a way that the LLM's text reasoning is misaligned. In late fusion, separate unimodal predictions are combined, which may allow an adversary to manipulate one modality's output to override the others. A compromised diagnosis from the image branch could dominate the final decision even if the text and lab values are correct.

Cross-modal attacks can be categorized based on the adversary's knowledge and access. White-box attacks assume full knowledge of the model parameters and gradients, enabling gradient-based optimization of perturbations across modalities [10]. Black-box attacks rely on querying the model or using transferable perturbations from surrogate models. In clinical settings, white-box attacks are less likely but still possible if model weights are leaked or if the adversary has insider access. More concerning are black-box attacks that exploit the model's public API, as many clinical LLM agents are deployed as cloud services. Additionally, physical-world attacks can occur, such as manipulating a patient's medical image by adding benign-appearing artifacts or altering text in a referral letter.

The clinical consequences of successful cross-modal attacks range from delayed diagnoses to incorrect treatment recommendations. For instance, an adversary could modify a digit in a laboratory result to simulate a life-threatening condition, causing the LLM agent to recommend unnecessary interventions. Alternatively, an attack could suppress a genuine abnormality by introducing conflicting signals in another modality, leading to missed detection. The cascade effect in LLM reasoning amplifies these risks, as the model may generate plausible but false explanations that convince clinicians to override their own judgment.

4. Adversarial Defense Architectures

Designing effective cross-modal defenses requires a multi-layered architectural approach that addresses both pre-processing and in-model mechanisms. The first line of defense is input sanitization, which involves detecting and neutralizing potential adversarial perturbations before they reach the model. For medical images, techniques such as JPEG compression, median filtering, and wavelet denoising can remove high-frequency perturbations while preserving diagnostic features [11]. However, these methods may degrade subtle pathological patterns, especially in low-contrast images. Similarly, for text, spell-checking, paraphrasing, and embedding obfuscation can mitigate certain attacks but may alter clinical meaning. The trade-off between defense robustness and diagnostic fidelity must be calibrated per modality and clinical context.

A more sophisticated approach involves cross-modal consistency verification. The idea is to compare the outputs or intermediate representations of different modalities to detect inconsistencies indicative of an attack. For example, if the image encoder predicts a certain finding (e.g., pneumonia), the textual interpretation should be semantically consistent. Discrepancies beyond a learned threshold can trigger a human review or an alternative model. This method requires a well-calibrated consistency model trained on benign data and validated across diverse patient populations. Challenges arise because genuine medical inconsistencies do occur—for instance, a patient may have typical symptoms of pneumonia yet a clear chest X-ray due to early infection. Thus, consistency verification must incorporate uncertainty quantification and domain knowledge to avoid false positives.

Robust training methods, including adversarial training, have been extended to multimodal settings. In adversarial training, the model is trained on both clean and adversarially perturbed examples to learn invariant features. For cross-modal scenarios, one can generate perturbations across all modalities simultaneously or randomly perturb a subset. Some works have proposed modality-specific adversarial training with a shared classifier to improve robustness to cross-modal attacks [12]. However, the computational cost of adversarial training grows with the number of modalities and the size of the LLM. Moreover, there is evidence that adversarial training can reduce standard accuracy, which is particularly problematic in medical applications where false negatives are dangerous.

Another architectural avenue is the use of disentangled representations, where each modality is encoded into a latent space with explicit invariance to other modalities. This can be achieved through variational autoencoders with mutual information minimization or contrastive learning objectives [13]. When an adversarial perturbation is applied to one modality, the disentangled representation should remain stable for the other modality, enabling detection and correction. The challenge lies in ensuring that the disentanglement does not discard clinically relevant cross-modal correlations.

A promising direction is the integration of real-time anomaly detection modules that monitor the model's internal activations and output distributions. Unsupervised anomaly detection, such as one-class support vector machines or autoencoder-based reconstruction error, can flag inputs that deviate from the training distribution [14]. This aligns with the recent work on security enhancement for adversarial robust LLM agents in medical decision-making tasks [14], which advocates for dynamic monitoring and adaptive control. The anomaly detection system should be trained on a large corpus of benign clinical data and continuously updated as new variants of attacks emerge. However, false alarms could erode clinician trust, so thresholds must be carefully set and explainability mechanisms integrated.

5. Structural Trade-offs and Governance

Every defense mechanism entails trade-offs among robustness, accuracy, latency, interpretability, and fairness. For instance, input sanitization may reduce the model's ability to detect subtle findings, leading to a trade-off between adversarial robustness and clinical sensitivity. Cross-modal consistency verification adds latency because it requires separate forward passes and comparison logic, which may be unacceptable in time-critical settings like emergency triage. Robust training often degrades performance on clean data, necessitating a careful evaluation of the cost-benefit ratio for each clinical scenario. Governance structures must formalize these trade-offs through risk assessment frameworks that involve clinicians, data scientists, ethicists, and regulatory experts.

One governance approach is to implement tiered defense configurations based on the criticality of the decision. For low-risk tasks such as appointment scheduling or documentation assistance, lighter defenses may suffice. For high-risk tasks such as diagnosis or treatment recommendation, full multi-layered defenses with human-in-the-loop verification should be mandatory. This tiered system requires clear definitions of risk categories and periodic audits to ensure compliance. Additionally, governance must address the lifecycle of defenses: models are continually updated with new training data, and adversarial attacks evolve. A governance board should oversee the frequency of defense updates, the validation of new defense mechanisms on diverse datasets, and the management of false positive rates.

Institutional policies should also mandate transparency about the presence and nature of defenses. Clinicians must be informed when an LLM agent's output has been flagged as potentially adversarial, and they should have the ability to override the system's recommendation. This human oversight is a critical safety net, but it requires clinician training and system usability. Furthermore, governance must consider liability: if a defense fails and a patient is harmed, who is responsible? Clear legal frameworks are needed to allocate accountability among model developers, healthcare institutions, and regulatory bodies.

6. Deployment and Infrastructure Considerations

Deploying cross-modal adversarial defenses in real-world clinical environments imposes stringent infrastructure requirements. Medical LLM agents often operate on cloud platforms or on-premises servers with low-latency constraints. Adding defense modules increases computational demand, which may necessitate additional hardware or optimized inference pipelines. For example, running an anomaly detection model alongside the LLM agent may double the inference time unless the detection model is lightweight or parallelized. In resource-constrained settings, such as rural clinics with limited bandwidth or older hardware, lightweight defenses are essential. Edge deployment, where the LLM agent runs locally on a device, reduces latency but limits the complexity of defense models. A hybrid approach—running a simplified defense on the edge and a more thorough analysis in the cloud—could balance speed and robustness.

Data governance is another critical infrastructure component. Defenses that rely on consistency verification or anomaly detection require access to representative training data that covers the expected distribution of inputs. However, medical data are highly sensitive, and sharing across institutions is often restricted by privacy regulations such as HIPAA. Federated learning could enable defense model training without centralizing data, but it introduces additional attack surfaces. Moreover, the infrastructure must support continuous logging of input-output pairs for post-hoc analysis of suspected attacks. Such logs must be stored securely and anonymized to protect patient privacy.

Sustainability of defenses over time requires automated pipelines for retraining models on new benign and adversarial examples. This pipeline should incorporate active learning to identify adversarial examples encountered in the field, and should be integrated with the model update cycle of the LLM agent itself. The energy consumption of frequent retraining and inference for defense modules is a sustainability concern, particularly as healthcare systems aim to reduce their carbon footprint. Efficient model compression and distillation techniques can mitigate this, but they may also degrade defense effectiveness.

7. Robustness, Fairness, and Policy Implications

Adversarial defenses must be evaluated not only for their technical efficacy but also for their impact on fairness across patient demographics. Research has shown that adversarial training can amplify existing biases, as the perturbations used during training may correlate with group membership [15]. For example, if adversarial examples are generated based on image intensity patterns that differ across ethnicities, the model may become less accurate for certain groups even on clean inputs. Similarly, input sanitization methods like filtering may disproportionately affect images from patients with comorbidities or rare anatomies. Cross-modal consistency verification may also be biased if the consistency model is trained primarily on data from majority populations, leading to higher false positive rates for minority groups.

To address these concerns, fairness-aware adversarial defense should be an integral component of the design process. This involves collecting stratified demographic data, auditing defense performance across subgroups, and adjusting thresholds or defenses to minimize disparities. Regulation such as the European Union's AI Act and the U.S. FDA's proposed framework for AI/ML-based medical devices emphasize the need for fairness and transparency [16]. Future policy should explicitly require that cross-modal defenses be tested for equitable performance, potentially including adversarial robustness as a fairness metric.

Policy implications extend to international harmonization of standards. Because medical LLM agents are increasingly deployed across borders, differing regulatory requirements for adversarial defense could create barriers to innovation while also offering opportunities for best-practice sharing. International bodies like the World Health Organization could play a role in setting baseline requirements for cross-modal defense in CDSS. Additionally, policies must address the dual-use nature of defense research: techniques developed for defense can be repurposed for attack. Responsible disclosure and controlled sharing of findings are necessary.

Finally, the long-term goal is to build trust in medical LLM agents among clinicians, patients, and regulators. Robust cross-modal defenses are a necessary but insufficient condition for trust. Transparency about model limitations, explainable decision-making, and continuous post-market surveillance are equally important. The interplay between technical robustness and socio-technical trust will shape the trajectory of AI in healthcare for years to come.

8. Conclusion

Cross-modal adversarial defense for medical LLM agents in clinical decision support systems is a multifaceted challenge that requires a coordinated approach spanning technical architecture, governance, deployment infrastructure, and policy. This paper has presented a comprehensive framework that identifies the unique vulnerabilities of multimodal LLM agents, analyzes the trade-offs inherent in various defense mechanisms, and discusses the socio-technical implications of deploying such defenses in real-world clinical settings. We argued that no single defense suffices; instead, a multi-layered strategy combining input

sanitization, cross-modal consistency verification, robust training, and real-time anomaly detection is necessary. Governance structures must formally manage trade-offs among robustness, accuracy, latency, and fairness, while infrastructure must be designed for scalability, sustainability, and data privacy. Future research should focus on developing fairness-aware adversarial defenses, reducing computational overhead, and integrating human oversight with automated safeguards. As medical AI continues to mature, ensuring that these systems are resilient to adversarial manipulation is not merely a technical objective but an ethical imperative.

References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).
2. Xu, K., Zhang, S., Tang, H., & Lu, C. (2020). Adversarial attacks and defenses in multimodal deep learning: A survey. ACM Computing Surveys, 53(4), 1–38.
3. 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.
4. Bagdasaryan, E., & Shmatikov, V. (2022). Spying on your dog: Adversarial attacks and defenses in multimodal models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
5. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR).
6. Tran, A., & Søgaard, A. (2021). Multimodal consistency verification for adversarial robustness. In Proceedings of the European Conference on Computer Vision (ECCV).
7. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.
8. Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO).
9. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.
10. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy.
11. Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. In International Conference on Learning Representations (ICLR).
12. Zhao, P., Liu, S., & Tao, D. (2021). Multimodal adversarial training: A unified framework. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 4567–4582.
13. Huang, J., & Feng, J. (2022). Disentangled representation learning for robust multimodal recognition. In Advances in Neural Information Processing Systems (NeurIPS).

14. 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.

15. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NeurIPS).

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

17. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

18. Chen, P. Y., Sharma, Y., Zhang, H., Yi, J., & Hsieh, C. J. (2018). EAD: Elastic-net attacks to deep neural networks via adversarial examples. In AAAI Conference on Artificial Intelligence.

19. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.

20. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Innovations in Theoretical Computer Science (ITCS).