Multimodal Adversarial Defense Framework for Vision-Language Medical Agents in Intelligent Diagnostic Environments

Fernando L. Norris
School of Computing, Clemson University, Clemson, SC, USA.
fernando1983@clemson.edu

Hugo C. Garrett
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
hugo.garrett725@ku.edu

Davide Tiaz
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
diaz1982@missouri.edu

Abstract

The increasing deployment of vision-language models in clinical workflows has given rise to a new class of intelligent medical agents capable of jointly interpreting radiological images, electronic health records, and natural language queries. Despite their diagnostic promise, these multimodal systems inherit profound adversarial vulnerabilities that threaten patient safety, diagnostic equity, and institutional trust. This paper presents a comprehensive adversarial defense framework designed explicitly for vision-language medical agents operating in intelligent diagnostic environments. We reconceptualize robustness not as a post hoc patch but as a first-class architectural property spanning data ingestion, cross-modal alignment, reasoning transparency, and runtime governance. The framework integrates multi-layered defense strategies, including modality-specific sanitizers, cross-modal consistency verification, structured output constraints, and a policy enforcement layer grounded in regulatory standards. We examine structural trade-offs between detection latency and diagnostic throughput, explore fairness implications under adversarial perturbations that disproportionately affect underrepresented patient populations, and analyze sustainability concerns arising from continuous adversarial retraining. The discussion extends to deployment architectures across hospital edge servers and centralized cloud platforms, highlighting governance requirements for software as a medical device. Through cross-domain comparisons with autonomous vehicle perception pipelines and financial fraud detection systems, we distill lessons on fail-safe design and explainability. The proposed framework is not tied to a single model architecture but serves as a system-level blueprint for crafting resilient medical agents that maintain clinical accuracy and ethical integrity under evolving threat models. We conclude with a roadmap for regulatory co-design, continuous certification, and federated adversarial monitoring across healthcare institutions.

Keywords

multimodal adversarial defense, vision-language medical agents, diagnostic robustness, medical AI security, clinical decision support, trustworthy AI governance, intelligent healthcare infrastructure.

1. Introduction

The convergence of large-scale vision-language pre-training and domain-specific medical knowledge has enabled an unprecedented class of intelligent diagnostic agents capable of simultaneously processing chest radiographs, gross pathology images, and unstructured clinical text. These agents promise to augment radiologists, support clinical decision-making in resource-constrained settings, and reduce diagnostic latency in acute care scenarios [1, 2]. Architecturally, such systems rest on multimodal transformer backbones that embed visual and textual modalities into a shared representational space, enabling cross-modal reasoning that goes far beyond the siloed analysis typical of earlier medical artificial intelligence [3]. Yet the very sophistication that makes these systems powerful also expands their attack surface in ways that remain poorly understood and insufficiently addressed by the broader machine learning security community. A single adversarial perturbation injected into either the imaging pipeline or the natural language input can cascade through the joint embedding space, corrupting attention maps, altering differential diagnoses, and ultimately producing clinically dangerous outputs that appear superficially plausible.

The adversarial robustness of medical agents cannot be treated as a narrow technical problem confined to training-time data augmentation or input denoising. Instead, it must be understood as a system-level challenge that interweaves model architecture, data governance, clinical workflow integration, and regulatory accountability. Prior work in adversarial machine learning has primarily focused on single-modality settings, such as pixel-level perturbations of natural images recognized by convolutional networks [4, 5]. In the medical domain, researchers have demonstrated that subtle adversarial modifications to retinal fundus photographs or dermatoscopic images can cause misclassification by deep learning models trained for disease detection [6, 7]. While alarming, these studies do not capture the complex failure modes that arise when a model must reconcile a perturbed chest CT with a subtly altered clinical history note that references the same anatomical region. In such contexts, the adversarial objective is not merely to fool an image classifier but to manipulate a dialog-capable agent into generating a coherent but erroneous report, potentially influencing downstream treatment decisions.

This paper introduces a multimodal adversarial defense framework purpose-built for vision-language medical agents in intelligent diagnostic environments. Rather than proposing a new attack or a single algorithmic countermeasure, we articulate a layered defensive architecture informed by lessons from critical infrastructure protection, safety-critical cyber-physical systems, and regulatory science. The framework repositions adversarial robustness as an architectural invariant that must be embedded in data preprocessing, cross-modal fusion, output generation, and continuous deployment monitoring. Throughout the discussion, we emphasize structural trade-offs—between detection sensitivity and clinical throughput, between fairness across demographic groups and uniform perturbation suppression, and between the computational sustainability of defense mechanisms and the carbon footprint of constant retraining. By drawing on illustrative case studies and cross-domain analogies, we aim to provide a foundational reference for system designers, policy-makers, and clinical informaticians charged with deploying the next generation of safe medical artificial intelligence.

2. The Multimodal Threat Landscape in Medical Diagnosis

A principled defense framework must begin with a clear-eyed assessment of the threat landscape. Medical diagnostic environments present unique adversarial constraints that distinguish them from open-domain image captioning or visual question answering. First, the cost of a false negative or false positive is measured in patient harm rather than inaccuracy on a leaderboard. Second, adversaries may possess partial knowledge of the model's training distribution and institutional protocols, especially in contexts where open-source foundation models serve as the backbone and only lightweight fine-tuning adapts them to local clinical data. Third, the attack may target any point in the multimodal chain: the image acquisition device, the digital imaging and communications in medicine pipeline, the natural language preprocessing layer that tokenizes referral notes, or the alignment module that projects visual features into the language model's embedding space. This heterogeneity necessitates a defense posture that is itself multimodal, context-aware, and tightly coupled to the semantics of the clinical task.

Consider a radiologist assistance agent that receives a frontal chest X-ray and a brief clinical indication such as "cough and fever for three days." An adversary could inject imperceptible perturbations into the image that cause the visual encoder to mislocalize an opacity, while simultaneously crafting a natural language adversarial example that subtly rewrites the indication to suggest a chronic condition. Even if each perturbation alone might not alter the final diagnosis, the joint effect can shift the agent's attention toward a false region of interest, producing a diagnosis of interstitial lung disease instead of pneumonia. Empirical studies on medical deep learning systems have shown that adversarial attacks can achieve high success rates while maintaining visual indistinguishability, and such attacks have been demonstrated on multiple imaging modalities including histopathology and fundus photography [7]. Extending these findings to the multimodal setting reveals that the cross-modal fusion mechanism, often implemented via cross-attention layers, can amplify subtle corruptions because the model learns to rely on spurious consistencies between perturbed inputs.

The threat actor model in a clinical context warrants careful delineation. While nation-state adversaries targeting medical artificial intelligence remain a speculative concern, more plausible threats include financially motivated actors seeking to influence insurance claim decisions, internal staff with technical knowledge who could tamper with local training data, and even unintentional adversarial drifts caused by software updates that alter preprocessing pipelines. Moreover, the increasing use of retrieval-augmented generation in medical agents introduces an additional vector: a compromised clinical knowledge base could inject misleading excerpts that coerce the agent's output even when both image and query are benign. Defenses that rely solely on training-time adversarial training [8] are insufficient for such dynamic threats because they cannot anticipate all novel attack compositions that emerge at the intersection of retrieval, vision, and generation.

3. Architectural Foundations of Vision-Language Medical Agents

To ground the defense framework, it is essential to characterize the typical architecture of a contemporary vision-language medical agent. Drawing from advances in general multimodal learning [3, 9] and their medical adaptations [10], such agents consist of several interconnected modules. A visual encoder, often a vision transformer pretrained on large-scale medical image corpora, produces a sequence of patch-level feature representations. Concurrently, a clinical language encoder processes the textual query and any retrieved context through a transformer stack fine-tuned on biomedical literature and electronic health

record narratives. These two streams are fused via a cross-modal alignment module that may employ contrastive learning objectives to bind semantically related visual and textual regions, forming a joint representation that is then consumed by a large language model decoder. The decoder generates free-text radiology reports, answers structured yes-no diagnostic questions, or produces structured findings in a templated format.

The architectural modularity of such systems offers both opportunities and vulnerabilities for defense. The visual encoder may be susceptible to adversarial patches specifically designed to exploit its inductive biases, while the language encoder can be undermined by carefully crafted token substitutions that preserve surface fluency but alter medical meaning. The fusion layer, often seen as the locus of robustness because it reconciles multiple evidence streams, can in fact become an adversary's amplifier if the alignment training does not explicitly penalize inconsistent co-occurrences. For instance, contrastive objectives that push matched image-text pairs closer while pushing mismatched ones apart can be subverted if an attacker generates a perturbed image that aligns better with a malicious text target than with the original report, effectively hijacking the agent's reasoning pathway. Recent medical visual question answering models have exhibited strong performance on curated benchmarks but have also shown brittleness when subjected to distribution shifts that are far less extreme than adversarial perturbations, indicating that the representations learned are not intrinsically robust to even mild input corruptions [10].

The decoder's role is equally critical. In a dialogue-oriented medical agent, the decoder must maintain factual consistency with both the visual evidence and the clinical context. However, standard language modeling objectives do not distinguish between faithful summarization and confident hallucination absent explicit constraints. An adversary who understands the decoder's tendency to generate fluent continuations can exploit this to induce a false finding that cascades through subsequent conversational turns, a phenomenon analogous to hallucination snowballing observed in retrieval-augmented models. Consequently, any viable defense must instrument not only the input layers but also the fusion and generation stages, imposing structural guardrails that limit the degrees of freedom available to an attacker while preserving clinical expressiveness.

4. Adversarial Vulnerabilities in Multimodal Medical Reasoning

Detailed examination of attack surfaces reveals three distinct classes of vulnerabilities that are particularly pernicious in medical reasoning: cross-modal consistency attacks, temporal drift manipulation, and adversarial context injection. A cross-modal consistency attack simultaneously perturbs image and text inputs such that each perturbed modality independently remains within the expected distribution of its encoder but the joint representation violates anatomical or pathological plausibility. For example, a chest X-ray might be modified to exhibit a subtle pleural effusion in the right costophrenic angle while the textual query is altered to include the phrase "blunting of the left costophrenic angle," creating a spatial contradiction that a human radiologist would instantly flag but that the fusion module, lacking explicit contradiction detection, might resolve by averaging attention weights and generating an amalgamated description of bilateral effusions. This type of attack underscores the need for consistency verification mechanisms that operate directly on the fused representation rather than on cleaned inputs alone.

Temporal drift manipulation is a less studied but equally dangerous threat. Medical agents that are deployed in dynamic clinical settings may receive inputs that evolve over time due to software updates, imaging device calibration shifts, or changes in clinical documentation

patterns. An adversary could introduce subtle perturbations that mimic this natural drift, gradually steering the model's output distribution toward a targeted error regime without triggering any single anomaly detector. Such slow-roll attacks are notoriously difficult to distinguish from benign domain shift and can persist for extended periods, eroding diagnostic accuracy across thousands of patient encounters. This vulnerability argues for defense architectures that incorporate runtime distributional monitoring and that can trigger automatic fallback to a validated static baseline when the live input distribution deviates beyond a statistically defined envelope.

Adversarial context injection targets the growing trend of equipping medical agents with retrieval components that pull relevant passages from clinical guidelines, PubMed abstracts, or local hospital protocols. An attacker who can compromise a small subset of the retrieval corpus can plant misleading documents that, when surfaced by the retriever for a common query, contaminate the reasoning chain. Because clinicians are more likely to trust outputs accompanied by seemingly authoritative citations, the potential for harm is magnified. Defenses against such threats must extend beyond conventional input sanitization to include trust provenance tracking and reference-level fact-checking. Recent work has proposed security enhancement methods tailored to large language model agents engaged in medical decision-making tasks, emphasizing the need for robust integration of external knowledge sources [11]. However, these methods have largely been developed in text-only settings, and their adaptation to vision-language pipelines remains an open challenge that this framework seeks to address structurally.

5. Multimodal Adversarial Defense Framework Design Principles

The proposed defense framework is organized around five core design principles: stratified sanitization, cross-modal verification, constrained decoding, runtime differential monitoring, and policy-informed escalation. Stratified sanitization acknowledges that adversarial perturbations manifest differently across modalities and must be countered with modality-appropriate techniques. For the visual stream, this entails a combination of pixel-level stochastic transformations, such as randomized spatial smoothing and frequency-domain filtering, alongside deep learning-based reconstruction modules trained specifically to remove medically implausible artifacts while preserving diagnostically relevant features. Importantly, these visual sanitizers must be calibrated on representative clinical image distributions rather than generic natural images, because aggressive smoothing can obliterate microcalcifications in mammograms or subtle interstitial changes that are clinically meaningful. The sanitization module thus operates under a tunable conservatism parameter that allows clinicians or system administrators to trade off between false alarm rate and acceptable residual perturbation.

For the textual stream, stratified sanitization includes syntax-aware back-translation, clinical entity grounding, and contradiction detection against a structured knowledge base of normal physiological ranges and anatomical compatibilities. A clinical note that mentions a "fracture of the left femur" when the corresponding pelvic X-ray has been perturbed to suppress the appearance of a fracture line would be flagged not merely because of lexical incongruity but because the structured extraction of anatomical landmarks fails to register an expected relationship. This integration of structured domain knowledge into the defense pipeline mirrors the dual-process reasoning recommended in clinical decision support systems, where intuitive pattern recognition is systematically checked against analytic rule-based verification.

Cross-modal verification constitutes the second pillar of the framework. After each modality has been independently sanitized, a dedicated verification module projects the cleaned

representations into a common semantic space and computes inconsistency scores using both learned and rule-based comparators. The learned comparators are trained on diverse counterfactual pairs that capture realistic clinical contradictions, including laterality mismatches, size discrepancies, and temporal impossibility. The verification module outputs an inconsistency map rather than a binary flag, allowing downstream components to determine whether the detected inconsistency is sufficient to trigger a rejection or whether the system can proceed with a confidence-qualified output. This graduated response is critical in clinical settings where an overzealous defense could cause diagnostic delays.

Constrained decoding imposes structural limitations on the autoregressive generation process of the language model decoder. Instead of free-form generation, the decoder is required to produce outputs that conform to a predefined clinical schema, such as a structured radiology report that includes separate sections for findings, impression, and recommendations. This schema acts as a scaffold that limits the space of plausible adversarial outputs because any generated text that deviates from the required structure is automatically suppressed. Constrained decoding can be implemented through token-level masking, reinforcement learning with grammar rewards, or post-hoc parsing and rejection. While this introduces a modest overhead and may reduce stylistic fluency, the clinical benefit of guaranteed structural integrity outweighs the cost, particularly in high-stakes scenarios such as emergency radiology.

Runtime differential monitoring provides continuous surveillance of the deployed agent by comparing its outputs against those of a lightweight frozen baseline model, which is never updated online and serves as a behavioral anchor. Significant divergence in diagnostic conclusions between the primary agent and the anchor, beyond what can be explained by benign calibration shift, triggers an alert and initiates a cascade of defensive actions ranging from logging for retrospective audit to automatic fallback to the baseline. This dual-model deployment strategy draws inspiration from the diversity-based fault tolerance techniques used in avionics systems, where multiple independent software implementations vote on safety-critical decisions. Finally, policy-informed escalation connects the technical defense mechanisms to the organizational and regulatory context. When a potential adversarial event is detected, the framework consults a policy engine that encodes institutional protocols, Food and Drug Administration guidance on software modifications, and data privacy regulations to determine the appropriate response, including notifying the clinical informatics team, quarantining the input for forensic analysis, and triggering model rollback in a compliant manner.

6. System-Level Integration and Deployment Considerations

Realizing the proposed framework in a functioning hospital environment demands attention to infrastructure architecture, latency budgets, and interoperability with existing health information technology ecosystems. A medical agent operating in an emergency department must return diagnostic impressions within tens of seconds, a constraint that conflicts with the computational expense of multimodal sanitization and verification. One architectural solution is to deploy the defense pipeline on a local inference server within the hospital’s secure network, allowing low-latency processing of sensitive patient data without requiring data to traverse wide area networks. The sanitization and verification modules can be accelerated using model distillation and quantization techniques that compress their computational footprint while preserving critical detection sensitivity. Edge deployment also facilitates compliance with privacy regulations that restrict the off-site transfer of protected health

information, an increasingly important consideration as regulatory agencies intensify their scrutiny of cloud-based artificial intelligence in healthcare.

Larger institutions may opt for a hybrid architecture in which the heavy-lifting components of the agent, including the large language model decoder, run on a private cloud with dedicated hardware accelerators, while the lightweight defense modules and the baseline anchor model execute at the edge. This topology introduces a dependency on the reliability and security of the network link, which must itself be protected against man-in-the-middle attacks that could substitute perturbed inputs in transit. Defense in depth, in this context, requires that the communication protocol employs mutual authentication and that the defense framework verifies the integrity of the received model responses using cryptographic signatures generated by the cloud inference endpoint. Such measures, while adding complexity, are standard practice in sectors such as financial transaction processing and are gradually being adopted in medical device interoperability standards.

The integration of the defense framework into health record systems and picture archiving and communication systems necessitates the development of a standardized interface that exposes the agent's confidence scores, inconsistency maps, and defensive actions to downstream clinical decision support modules. A radiologist reviewing an agent's flagged finding should see not only the agent's textual impression but also an indication of whether any cross-modal inconsistencies were detected and whether the output was generated under constrained decoding. This transparency can bolster clinician trust and enable more informed override decisions. From a software as a medical device perspective, any adaptive aspect of the defense framework, such as dynamic threshold adjustment based on deployment drift, may be subject to premarket review or postmarket surveillance requirements [12]. Consequently, the framework must be designed with an immutable core whose behavior has been validated in a predetermined operating envelope, while the adaptive components are externally configurable within pre-approved bounds. This separation of concerns simplifies regulatory compliance and aligns with emerging frameworks for change control in continuously learning medical systems.

7. Governance, Fairness, and Ethical Implications

Adversarial defenses in medical artificial intelligence are inseparable from questions of fairness and justice. A defense strategy that applies uniform perturbation filtering across all patient subgroups may inadvertently amplify disparities if the sanitization pipeline is less effective on images with higher melanin content, poorer contrast due to suboptimal acquisition conditions in under-resourced clinics, or rare pathological presentations that lie far from the training distribution of the sanitizer. The medical imaging literature has documented how deep learning models can exhibit performance gaps across demographic groups and imaging device manufacturers [13]. When adversarial defense mechanisms compound these pre-existing biases, they risk creating a two-tiered diagnostic system in which protected patient populations not only experience lower baseline accuracy but are also disproportionately subjected to adversarial rejection events that force manual override, increasing already burdensome workloads on clinicians in safety-net hospitals.

Mitigating fairness concerns within the defense framework requires explicit incorporation of subgroup-specific robustness metrics during the design and evaluation phases. The stratified sanitization modules must be trained on diverse image-text pairs that span the full demographic spectrum and must be validated using adversarial perturbations crafted against intersectional subgroups, not merely the aggregate population. Cross-modal verification

thresholds can be set differentially based on the known base rate of certain conditions in specific populations, but such adjustments must be transparent, clinically justified, and regularly audited to prevent the introduction of new disparities. Additionally, the policy-informed escalation engine should record demographic attributes associated with each defensive action, enabling post-hoc fairness audits that can be reviewed by institutional ethics committees and, where required, by regulatory bodies.

The broader governance implications extend to liability, accountability, and the evolving landscape of artificial intelligence regulation in medicine. If an adversarial attack successfully deceives a medical agent despite the presence of the defense framework, the allocation of responsibility among the model developer, the healthcare provider, the defensive module vendor, and the cloud infrastructure operator becomes contested. This ambiguity may chill adoption unless clear safe-harbor provisions are established, analogous to those enacted for medical device interoperability. The defense framework can contribute to accountability by generating an adversarial event log that captures the sanitized inputs, the inconsistency scores, the voting records among monitors, and the final action taken. Such logs, when combined with secure timestamping and cryptographic integrity, provide the evidential basis for root-cause analysis and continuous improvement. They also serve as a foundation for algorithmic auditing practices that are increasingly mandated by proposed artificial intelligence accountability legislation [14]. Furthermore, the defense framework's design should align with the ethical principle of explainability, ensuring that not only are adversarial inputs rejected but that the reason for rejection is communicated in clinically interpretable language, enabling a seamless handoff to human decision-makers.

8. Evaluation Strategies and Long-Term Sustainability

Evaluating a multimodal adversarial defense framework poses distinct methodological challenges that go beyond standard metrics of attack success rate and clean accuracy retention. In diagnostic settings, clinical equivalence studies are necessary to demonstrate that the instrumented agent performs no worse than the undefended version on a representative sample of unperturbed real-world cases, a requirement that resonates with the statistical paradigms of clinical trials. Adaptive red-teaming, in which a panel of security researchers and clinicians collaboratively design novel attack strategies that exploit domain-specific knowledge, provides a more realistic stress test than automated evaluation with gradient-based perturbations. The framework should be subjected to longitudinal testing in a simulated clinical environment where attack patterns evolve over time, mirroring the adversarial dynamics of real-world deployment.

Sustainability considerations are equally pressing. Adversarial training and periodic retraining of defense modules consume substantial computational resources, contributing to the carbon footprint of the overall system [15]. The framework's design philosophy of using lightweight, frozen baseline models for monitoring and constrained decoding reduces the need for frequent retraining of large language models, which are the dominant source of energy consumption. Moreover, the modular architecture allows individual sanitization and verification components to be updated independently without triggering a full validation cycle of the entire agent, a property that lowers the barrier to responsive defense against newly discovered attacks. The defense framework also incorporates capacity planning for federated adversarial monitoring, in which multiple hospitals contribute de-identified inconsistency logs to a shared threat intelligence repository without centralizing sensitive patient data. This federated approach, facilitated by privacy-preserving aggregation protocols, enables the healthcare

sector to collectively respond to emerging adversarial phenomena at a speed that no single institution could achieve alone, while respecting the data sovereignty constraints that govern cross-border health data flows.

9. Conclusion

The integration of vision-language models into medical diagnosis represents a transformative opportunity to enhance clinical accuracy, reduce workflow bottlenecks, and expand access to specialist expertise. However, these benefits are contingent upon the robustness of the underlying intelligent agents against adversarial manipulation that can arise from both malicious actors and unintended environmental shifts. This paper has articulated a comprehensive defense framework that reimagines adversarial robustness as a system-level property permeating data ingestion, multimodal fusion, generation, monitoring, and governance. By embedding stratified sanitization, cross-modal verification, constrained decoding, and policy-informed escalation into a cohesive architecture, the framework addresses the unique failure modes of multimodal medical reasoning while remaining cognizant of the clinical, regulatory, and ethical dimensions that distinguish healthcare from other application domains. The framework's emphasis on fairness, transparency, and sustainable deployment practices reflects a conviction that security and equity are mutually reinforcing design objectives. Future work should pursue empirical validation of the framework in partnership with clinical sites, develop standardized adversarial benchmarks for vision-language medical agents, and engage regulatory bodies to co-evolve approval pathways that accommodate the defensive adaptability required in an adversarial landscape. Ultimately, the safe clinical translation of medical artificial intelligence will depend not on any single algorithmic breakthrough but on the deliberate construction of resilient socio-technical systems that embed adversarial preparedness into their architectural DNA.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
2. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML).
4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).
5. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. ICLR Workshop.
6. 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.
7. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition, 110, 107633.
8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR).

9. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning (ICML).

10. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., ... & Ho, C. (2023). PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.09015.

11. 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.

12. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

14. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

15. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

16. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265.

17. Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

18. Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. Proceedings of the 36th International Conference on Machine Learning (ICML).

19. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI).

20. U.S. Food and Drug Administration. (2021). Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. https://www.fda.gov/media/145022/download

21. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (SP).

22. Wong, E., & Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. Proceedings of the 35th International Conference on Machine Learning (ICML).

23. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for

internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT).