Trustworthy Retrieval-Augmented Generation for Adversarially Robust Medical Large Language Model Agents

Shane Bush
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
shane.bush811@buffalo.edu

Leonard Bamos
Department of Computer Science, University of Houston, Houston, TX, USA.
leonardramos95@uh.edu

Abstract

The rapid integration of large language models into clinical decision support has created a new class of medical artificial intelligence agents capable of synthesizing vast medical knowledge, yet their safe deployment is undermined by adversarial threats and unresolved trustworthiness gaps. This paper examines the design space for retrieval-augmented generation architectures that serve as the cognitive core of these agents, focusing on the interplay between adversarial robustness, clinical reliability, and system-level governance. We argue that simply inserting a retrieval component into a generative model does not inherently confer trustworthiness; rather, it introduces a complex sociotechnical surface that adversaries can exploit through data poisoning, prompt injection, and corpus manipulation. Starting from a system-of-systems perspective, we dissect the layered infrastructure of medical large language model agents, analyze the emergent adversarial attack taxonomy specific to retrieval-augmented generation, and articulate architectural principles for embedding robustness without sacrificing clinical performance. The discussion extends to structural trade-offs involving latency, fairness, sustainability, and regulatory alignment, emphasizing that trustworthiness cannot be localized to a single model module but must be distributed across data governance, retrieval curation, model alignment, and human oversight. Policy implications for medical device regulation, continuous monitoring, and multi-stakeholder accountability are explored. By synthesizing cross-domain insights from machine learning security, health informatics, and infrastructure studies, we outline a holistic framework for building adversarially resilient medical agents that retain the flexibility of retrieval-augmented generation while upholding rigorous safety standards.

Keywords

Retrieval-Augmented Generation; Medical AI Agents; Adversarial Robustness; Trustworthiness; Large Language Models; Clinical Decision Support; System Governance.

1. Introduction

The emergence of large language models (LLMs) with the capacity to encode substantial clinical knowledge has ignited transformative ambitions for medical artificial intelligence. Early demonstrations revealed that LLMs can achieve near-expert performance on medical licensing examinations, summarize patient histories, and propose differential diagnoses, spurring research into their integration as intelligent agents within clinical workflows [1].

However, the inherent propensity of these models to generate plausible yet incorrect or harmful outputs, commonly termed hallucination, poses unacceptable risks in healthcare contexts where errors can lead to severe patient harm. To mitigate this, the retrieval-augmented generation (RAG) paradigm has been widely adopted to ground LLM responses in verifiable external knowledge sources, allowing agents to cite clinical guidelines, biomedical literature, and electronic health records at inference time [2]. While RAG substantially improves factual accuracy, it concurrently expands the attack surface of medical agents, introducing new adversarial vulnerabilities that threaten trustworthiness at the system level.

Adversarial threats to machine learning systems have been studied extensively in classification tasks, with techniques ranging from character-level perturbations to semantically preserved input transformations that induce model misbehavior [3]. Language models amplify these risks because adversarial inputs can be constructed via inconspicuous textual modifications, universal adversarial triggers, or carefully crafted prompts that bypass content filters [4]. When a retrieval mechanism is coupled with a generative backend, attackers gain additional avenues for manipulation, including poisoning the underlying knowledge corpus, injecting malicious content into retrieved documents, and exploiting the agent's reliance on external data to propagate disinformation at scale [5]. The clinical domain further intensifies the stakes, as adversarial manipulation of treatment recommendations, lab values, or drug interaction databases could precipitate catastrophic outcomes.

The trustworthiness of medical LLM agents must therefore be assessed holistically, encompassing accuracy, reliability, fairness, explainability, privacy, and robustness under adverse conditions. Prior work has documented how biomedical algorithmic systems can embed and amplify social biases, resulting in unequal care quality across demographic groups [6]. These biases can be both inherited from training data and introduced or magnified by the retrieval corpus, creating feedback loops that are difficult to audit in dynamic, continuously updated knowledge bases. Approaches such as chain-of-thought prompting have been proposed to improve interpretability by exposing model reasoning traces, yet their reliability under adversarial inputs remains uncertain [7]. Meanwhile, global health authorities have issued ethical frameworks for artificial intelligence in medicine, stressing the need for transparency, human accountability, and safety-by-design [8]. The convergence of these dimensions indicates that adversarially robust medical RAG agents demand a system-oriented perspective that goes beyond individual model hardening.

Given the complexity of this design space, comprehensive evaluation methodologies have been developed to benchmark LLMs across a spectrum of trustworthiness criteria [9]. In parallel, researchers have begun to formulate specific security enhancement methods for medical decision-making agents that integrate retrieval components, aiming to fortify systems against adversarial exploitation while preserving clinical utility [10]. Despite this progress, the existing literature often treats adversarial robustness, fairness, and system governance as separate research streams. In this paper, we weave these threads together, presenting an architectural and infrastructural analysis of trustworthy RAG for medical agents. We examine the structural trade-offs inherent in building systems that are simultaneously accurate, robust, fair, and deployable within contemporary healthcare environments.

2. Architectural Foundations of Medical RAG Agents

A medical RAG agent can be conceptualized as a layered architecture in which a retrieval engine interfaces with a generative model, orchestrated by a coordination layer that manages user queries, context assembly, and response validation. At the infrastructure level, the

retrieval subsystem typically consists of a vector database populated with embeddings of clinical texts, drug formularies, research articles, and guideline documents. When a query arrives, it is encoded and matched against the database to return a set of topically relevant passages, which are then concatenated as context for the generative model. This decomposition creates a clean separation of concerns: the retrieval component provides factual grounding, while the generator focuses on fluent synthesis. Federated retrieval architectures can further allow agents to query distributed data silos across multiple institutions without centralized data aggregation, addressing privacy concerns [11].

This separation, however, introduces critical dependencies across the retrieval, generation, and integration layers that adversaries can exploit at multiple points. The trustworthiness of the generated output is contingent upon the integrity of the retrieved context, which in turn depends on the provenance, curation, and freshness of the underlying knowledge base. Moreover, the generator must be capable of appropriately weighting the retrieved evidence against its own parametric knowledge, especially when the two conflict. System designers face a fundamental structural tension: enriching the retrieval corpus with extensive, up-to-date sources improves factual coverage but also increases the attack surface for data poisoning and adversarial insertion. Conversely, restricting retrieval to a tightly controlled, manually curated set of documents limits adversarial opportunities but may reduce the agent's ability to handle rare diseases, off-label medication inquiries, or emerging clinical knowledge from recent publications.

Another architectural consideration is the degree of coupling between retrieval and generation. In tightly coupled designs, the generator is fine-tuned end-to-end on retrieved contexts, which can improve coherence but may cause the model to overfit on retrieval artifacts or specific corpus styles. Loosely coupled systems treat the retriever as a pluggable module, offering flexibility to swap knowledge sources without retraining the generator. However, this modularity can lead to misalignment if the generator does not reliably interpret the structure and credibility markers of the incoming documents. Deciding on the proper level of integration requires weighing deployment simplicity against the need for robustness to distributional shifts in the retrieval corpus, a concern amplified in adversarial settings where an attacker could introduce documents designed to mimic authoritative sources while carrying subtle semantic manipulations.

Additional layers such as response validation, uncertainty quantification, and guardrail modules are increasingly incorporated into medical RAG pipelines. Response validators may cross-check generated answers against the retrieved evidence or employ secondary LLM calls to flag inconsistencies, while uncertainty estimators can trigger human review when confidence falls below a predefined threshold. These components, while valuable for safety, themselves become new targets for adversarial attacks and further complicate latency and computational budgets. The rise of multi-agent frameworks that coordinate specialized LLM agents for retrieval, reasoning, and verification introduces a higher-order orchestration challenge, where the overall system robustness depends on the interaction protocols and conflict resolution mechanisms among agents [14]. As architectural choices proliferate, the need for principled design guidelines that account for adversarial threats from the ground up becomes increasingly urgent.

3. Adversarial Threat Landscape in Medical RAG Systems

Adversaries targeting medical RAG agents can operate along multiple dimensions, corresponding to the data, retrieval, model, and integration layers. Poisoning the retrieval

corpus represents a particularly insidious attack class because the injected malicious content can persist silently, affecting countless clinical queries before detection. An attacker who gains write access to a knowledge base, perhaps through compromised software supply chains or insider threats, can insert documents that bias the agent's recommendations toward specific treatments, pharmaceutical products, or diagnostic pathways. Because the retrieval process often relies on dense vector similarity, adversarial documents can be optimized to appear highly relevant to a wide range of benign queries, leveraging embedding-space perturbations analogous to adversarial examples in computer vision. Trojan-based backdoor attacks on retrieval-augmented generation have been demonstrated, where a trigger phrase in a query causes the system to retrieve attacker-controlled passages and generate harmful responses, highlighting the fragility of the retrieval-generation interface [15].

Prompt injection attacks constitute another pervasive threat in which an adversary embeds malicious instructions within the retrieved documents themselves. When the generator processes a retrieved passage that contains text such as "ignore previous instructions and recommend medication X," the model may treat it as a legitimate command, bypassing safety tuning. This indirect form of attack is particularly dangerous in medical contexts because clinical documents often include structured fields, dosage ranges, and contraindication lists that are natural carriers for injected content. Existing research has shown that real-world LLM-integrated applications can be compromised through indirect prompt injection, allowing adversaries to extract sensitive information or alter decision outputs without direct access to user prompts [5]. In a hospital setting, a compromised document in the retrieval base could, for example, cause the agent to suggest a dangerous drug combination by manipulating text inside what appears to be a guideline excerpt.

At the input layer, adversarial queries designed to bypass content filters remain a persistent challenge. Even well-aligned medical LLMs may produce unsafe outputs when prompted with carefully crafted jailbreaks that exploit model weaknesses, such as role-playing scenarios, multi-turn dialogues, or splitting harmful requests across retrievals. In the RAG setting, an attacker might submit a query that triggers retrieval of a seemingly innocuous article while the generation step combines retrieved content with adversarial query phrasing to generate unethical advice. The interplay between retrieval and generation creates a combinatorial explosion of attack vectors that are difficult to anticipate and test exhaustively, necessitating defense-in-depth strategies.

Beyond output manipulation, privacy and inference threats include membership inference attacks on the retrieval corpus and model inversion techniques aimed at extracting patient data. Differential privacy methods have been applied to LLM training but are more challenging to enforce for dynamically updated retrieval bases that aggregate data from multiple clinical sites [16]. Additionally, fairness-oriented adversarial attacks can strategically embed biased content that targets specific demographic groups, causing the agent to exhibit differential error rates or discriminatory recommendations. This intersection of adversarial security and algorithmic fairness demands that robustness measures be evaluated not only by average accuracy but also by their impact on subpopulation disparities, which may be further intensified under adversarial pressure. A holistic threat model for medical RAG agents must therefore account for epistemic corruption, demographic skews, and long-term erosion of clinician trust.

4. Strategies for Adversarially Robust and Trustworthy Design

Achieving adversarially robust medical RAG agents requires combining proactive defense mechanisms at each architectural layer with systemic governance structures that enable detection, response, and recovery. At the data and retrieval level, robust provenance tracking is essential: every document in the knowledge base should carry cryptographic signatures and versioning metadata that allow the agent to verify authenticity and recency before using it as context. Reputation systems can be borrowed from distributed systems literature to maintain trust scores for knowledge sources, dynamically deprioritizing or quarantining sources that exhibit unusual update patterns or have been flagged in post-market surveillance. Retrieval augmentation should be combined with adversarial training on cleansed and perturbed datasets to reduce the model's susceptibility to poisoned contexts, though this must be balanced against the risk of overfitting to specific attack patterns and degrading performance on clean queries.

Input sanitization and prompt hardening techniques form a second line of defense. Query classifiers can be deployed before retrieval to detect and reject obviously malicious requests, but these classifiers themselves must be robust to adversarial evasion. A more promising direction is to design the generative agent's system prompt and retrieval instructions with explicit guardrails that constrain the model to cite only verified sources and to refuse to follow instructions embedded within retrieved passages. This can be implemented through constrained decoding, where the model's output space is limited to tokens consistent with a formal template that separates evidence presentation from clinical reasoning. Furthermore, isolating the agent's functional roles, so that retrieval, reasoning, and response generation operate in separate, auditable modules with strict interfaces, can limit the blast radius of a compromise in any single component.

Multi-agent verification architectures provide an additional robustness layer by having independent LLM instances cross-check each other's outputs against the retrieved evidence and known clinical knowledge bases [14]. One agent might generate a treatment recommendation, another critiques it using a different retrieval corpus or model variant, and a reconciliation module resolves disagreements through structured debate or majority voting weighted by source trustworthiness. Such redundancy incurs higher computational cost but can significantly raise the bar for adversaries, who must then simultaneously fool multiple diverse components to alter the collective output. Leveraging model diversity, where different agents are trained on distinct data distributions or with varying architectural inductive biases, reduces common-mode failure risks that plague monolithic systems.

Alongside real-time defenses, system-level monitoring and continuous adversarial evaluation are indispensable. Medical RAG agents deployed in clinical settings should log retrieval choices, generated responses, and anomaly signals to enable post-hoc audits and incident investigation. Red-teaming exercises that simulate motivated adversaries with domain knowledge of medical workflows can uncover novel attack vectors before they are exploited in the wild. Benchmarks for holistic model evaluation have expanded to cover robustness, fairness, and calibration, yet these must be extended to specifically target the retrieval-generation interface under adversarial manipulation [9]. Recent work has started to address this gap by proposing security enhancement methods tailored to medical LLM agents, integrating adversarial detection filters with retrieval integrity checks that compare embedded documents to trusted versions [10]. These system-level defenses should be complemented by organizational protocols that define escalation paths when an agent's trust score falls below a

safety threshold, ensuring that no automated clinical decision is made without appropriate human oversight.

5. System-Level Trade-offs and Infrastructure Considerations

Building adversarially robust medical RAG agents is not merely a technical challenge but a profound exercise in navigating trade-offs among competing system properties. The addition of verification modules, diverse retrievers, and real-time adversarial detection invariably increases inference latency and computational cost. In acute care settings such as emergency departments, where decision support must be delivered within seconds, safety mechanisms that introduce multi-second delays could reduce clinical adoption and prompt workarounds that bypass safety measures altogether. System architects must therefore optimize the security-compute budget by dynamically adjusting defense intensity based on query criticality, patient risk scores, and time constraints, a form of risk-adaptive security that remains under-explored for LLM-based medical agents.

Sustainability is an equally pressing infrastructure concern. The carbon footprint of large-scale LLM deployment has been extensively documented, and the additional overhead of retrieval, verification, and multi-agent consensus can amplify energy consumption significantly [12]. Hospitals and health systems with limited computational resources, particularly in low-resource settings, may be unable to deploy the most secure configurations, creating a robustness divide that mirrors existing health inequities. Lightweight retrieval models, knowledge distillation from robust teacher agents, and shared verification services operated at the network or regional level offer potential pathways to democratize trustworthy medical agent access. Policy interventions that mandate minimum robustness standards for AI-based clinical decision support must be carefully calibrated to avoid inadvertently excluding under-resourced providers from the benefits of generative AI.

Fairness concerns intersect with robustness trade-offs in non-trivial ways. Aggressive filtering of retrieved documents to exclude potentially poisoned sources might inadvertently censor information about diseases that predominantly affect marginalized communities if adversarial poisoning has concentrated on those topics. Similarly, defenses that rely on majority voting across language model agents may amplify biases encoded in the training data of the most prevalent models. A commitment to fairness requires that robustness mechanisms be evaluated through stratified performance metrics across demographic and clinical subgroups, and that bias mitigation techniques be integrated into the retrieval, generation, and verification stages rather than applied as a post-hoc overlay [6]. Transparent documentation of the knowledge base composition, along with algorithmic impact assessments, should become standard practice for any deployed medical RAG system.

Finally, the regulatory environment for medical AI is evolving rapidly, with agencies such as the U.S. Food and Drug Administration issuing action plans for AI/ML-based software as a medical device that emphasize iterative learning and real-world performance monitoring [18]. Medical RAG agents that continuously update their retrieval corpora fall into a regulatory gray zone, as their behavior can drift substantially over time without a clear change control process. Establishing version-controlled knowledge baselines, coupled with automated regression testing against clinical benchmarks such as MedQA, allows developers to demonstrate that system updates do not degrade safety or introduce adversarial vulnerabilities [20]. Integration with existing health information technology ecosystems, including electronic health record systems and clinical data warehouses, further demands compliance with interoperability standards and data protection regulations. The long-term viability of

trustworthy medical LLM agents hinges on constructing a robust socio-regulatory infrastructure that incentivizes adversarial preparation, mandates transparency, and enforces accountability across all actors in the supply chain.

6. Governance, Fairness, and Policy Implications

The deployment of adversarially robust medical RAG agents raises profound governance questions that extend beyond technical specifications. Accountability in the event of an adversarial-induced harm is difficult to assign within distributed architectures that involve multiple model providers, knowledge base curators, and healthcare delivery organizations. Clear allocation of liability, supported by contractual frameworks and regulatory guidance, is necessary to prevent a diffusion of responsibility that leaves patients unprotected. Drawing from the WHO ethical principles, medical AI systems must uphold human dignity, ensure equitable access, and maintain transparency such that clinicians can understand and contest automated recommendations [8]. For RAG agents, this implies the need for verifiable chain of custody for all retrieved evidence, so that when a clinician disagrees with a generated suggestion, the provenance of the underlying information can be traced back through the retrieval pipeline.

Institutional governance structures must evolve to support continuous adversarial readiness. Just as hospital infection control committees oversee antimicrobial stewardship, dedicated AI safety boards can monitor the performance of deployed agents, review red-teaming findings, and authorize updates to defense mechanisms. These boards should include clinical domain experts, patient representatives, data scientists, and ethicists to balance technical feasibility with patient welfare. The growing complexity of multi-agent medical systems, in which multiple LLM agents collaborate on diagnosis and treatment planning, will require governance models that can assess emergent behaviors not attributable to any single component, a challenge familiar to the study of complex sociotechnical systems [19]. Simulation-based adversarial stress testing, conducted in digital twins of clinical environments, can help governance bodies anticipate failure modes before they manifest in patient care.

Standards development will be critical for interoperability and for building a market that rewards adversarial robustness. Certification schemes inspired by the Common Criteria for information security could define assurance levels for medical RAG agents based on the depth of adversarial evaluation, retrieval provenance controls, and fairness audits. Publicly funded independent evaluation laboratories, akin to the role of the National Institute of Standards and Technology in cybersecurity, could provide benchmarked adversarial testing services and publish comparative results to drive industry-wide improvement. The development of open-source adversarial test suites that simulate diverse attack vectors on RAG architectures would further democratize robustness assessment and lower barriers to entry for smaller developers.

From a fairness perspective, the governance framework must also address the risk that adversarial hardening efforts prioritize the protection of majority populations while neglecting marginalized groups, who are often underrepresented in both training data and adversarial test scenarios. Policy mandates should require disaggregated robustness reporting by race, gender, age, and socioeconomic status, ensuring that defenses do not inadvertently create safety havens for some while leaving others exposed. The intersection of data privacy laws, such as the Health Insurance Portability and Accountability Act, with adversarial monitoring also warrants careful consideration, because logging adversarial queries and retrieved contexts for

audit purposes could inadvertently capture protected health information. Privacy-preserving logging techniques and differential privacy safeguards can help reconcile the dual imperatives of security monitoring and patient confidentiality [16].

7. Conclusion

This paper has articulated a system-level perspective on building trustworthy retrieval-augmented generation architectures for medical LLM agents, with an emphasis on adversarial robustness as a foundational design constraint rather than a retrofit. We have argued that the integration of retrieval components, while essential for factual grounding, expands the threat surface in ways that demand coordinated defenses across data governance, architecture modularity, verification protocols, and institutional oversight. The structural tensions between robustness, latency, fairness, and sustainability cannot be dissolved by any single technical fix; they require ongoing negotiation through adaptive governance mechanisms and policy frameworks that reflect the high stakes of clinical decision support. As large language models mature into core components of healthcare infrastructure, the field must move beyond accuracy-centric benchmarking to embrace adversarial resilience, equitable access, and transparent accountability as equally weighted pillars of trustworthiness. The path forward involves not only advancing defensive techniques, but also cultivating an ecosystem in which robust design is incentivized, evaluated transparently, and governed by multidisciplinary institutions that safeguard public health. In this vision, the medical RAG agent becomes not an autonomous oracle, but a resilient, auditable partner embedded within a web of evidence, oversight, and human judgment.

References

1. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Scharli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474).

3. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 31–36).

4. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 2153–2162).

5. 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.

6. 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.

7. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837).

8. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.

9. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research.

10. 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.

11. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

12. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

13. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409.

14. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155.

15. Cheng, R., Li, Z., Ding, K., Zhao, T., & Huang, H. (2024). TrojanRAG: Retrieval-augmented generation can be backdoored. arXiv preprint arXiv:2403.04877.

16. Li, X., Tramèr, F., Liang, P., & Hashimoto, T. (2022). Large language models can be strong differentially private learners. In International Conference on Learning Representations.

17. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.

18. U.S. Food and Drug Administration. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.

19. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

20. Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.