Knowledge Graph-Enhanced Secure Large Language Model Agents for Explainable Clinical Decision-Making under Adversarial Attacks

Rdrian Pichards
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
rdrianrichards95@oregonstate.edu

Jingzhong Yin
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
jingy@missouri.edu

Mikkel Koskinen
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
mikkelk@ucf.edu

Abstract

The integration of large language model (LLM)-based autonomous agents into clinical decision support systems promises transformative gains in diagnostic accuracy, treatment personalization, and workflow efficiency. However, the deployment of such agents in high-stakes medical environments is contingent upon resolving fundamental tensions among security, explainability, and reliability under adversarial conditions. This paper presents a systems-level architectural framework that couples knowledge graph-enhanced reasoning with multi-layered defense mechanisms to enable secure, interpretable, and adversarially robust LLM agents for clinical decision-making. We examine the structural trade-offs involved in embedding curated biomedical knowledge graphs as semantic scaffolding that constrains agent reasoning, mitigates hallucination, and provides verifiable provenance for generated recommendations. The architecture incorporates adversarial training, input sanitization layers, prompt integrity verification, and runtime monitoring as part of a defense-in-depth strategy against evolving threat vectors including prompt injection, semantic perturbation, and model extraction. Through a conceptual analysis grounded in cross-domain comparisons with critical infrastructure systems, we analyze how the interaction between symbolic knowledge structures and subsymbolic inference engines can simultaneously enhance explainability and resilience. We further address governance, fairness, sustainability, and policy implications arising from the deployment of such agents within regulated healthcare ecosystems. The discussion highlights the need for standardized evaluation protocols, continuous certification pipelines, and inclusive design practices that account for dataset shifts, demographic representation, and long-term operational viability. The paper advances a holistic perspective that treats security, interpretability, and clinical utility not as separable modules but as interdependent properties that must be co-engineered across the entire agent lifecycle.

Keywords

large language models, knowledge graphs, clinical decision support, adversarial robustness, explainable AI, multi-agent systems, healthcare security.

1. Introduction

The emergence of large language model (LLM)-based autonomous agents has reshaped the landscape of clinical informatics by enabling natural language interaction with complex medical knowledge bases, automated summarization of patient records, and real-time evidence synthesis [1], [2]. These agents, capable of tool use, memory-augmented reasoning, and multi-step planning, are increasingly seen not merely as assistive chatbots but as semi-autonomous decision-support entities that could participate in differential diagnosis, treatment recommendation, and medical documentation. However, the introduction of such powerful generative capabilities into safety-critical clinical workflows exposes healthcare systems to a new class of adversarial threats that can manipulate agent behavior through carefully crafted prompts, poisoned training data, or semantic backdoors [3]. The simultaneous demand for transparency in clinical reasoning further complicates the design space, as deep neural models remain largely opaque and susceptible to confidently generating plausible but medically incorrect outputs. Knowledge graphs have been advanced as a promising architectural component to anchor LLM outputs to structured, curated, and auditable biomedical knowledge, thereby reducing factual inconsistency and improving explainability [4]. Yet the interaction between knowledge graph constraints and adversarial robustness has received insufficient attention from a systems engineering perspective, particularly with respect to how such integrated architectures can be hardened against adaptive adversaries that exploit the interface between symbolic and neural components.

This paper addresses the gap by proposing a comprehensive architectural framework for knowledge graph-enhanced secure LLM agents tailored to clinical decision-making under adversarial conditions. Rather than focusing narrowly on a single defense technique or a particular model architecture, we adopt a systems-level lens that emphasizes structural trade-offs, layered defense strategies, governance requirements, and lifecycle sustainability. The complexity of clinical environments, characterized by heterogeneous data sources, stringent regulatory oversight, evolving medical knowledge, and high consequences for error, demands an approach that co-designs security, explainability, and clinical effectiveness from the earliest architectural decisions. The central argument is that knowledge graphs can serve not only as a knowledge retrieval substrate but also as a semantic validation layer that filters out adversarial perturbations, constrains the agent's generative space to clinically plausible trajectories, and furnishes human-readable justifications anchored in trusted sources. At the same time, the graph structures themselves become potential attack surfaces, necessitating integrity protections that span data ingestion, entity resolution, and real-time querying.

We structure the remainder of the paper as follows. Section 2 surveys the intersecting foundations of LLM agents, adversarial robustness in natural language processing, knowledge graphs in medicine, and explainable AI. Section 3 presents the core architectural framework, detailing how knowledge graph modules, adversarial defense pipelines, and explanation generators can be integrated. Section 4 analyzes the adversarial robustness of the proposed system under a defense-in-depth strategy, comparing different threat models and mitigation layers. Section 5 addresses explainability and trustworthiness from both technical and clinical adoption perspectives. Section 6 discusses infrastructure, deployment governance, fairness, and sustainability. Section 7 concludes with implications for future system design and regulatory pathways.

2. Background and Related Work

The evolution from standalone language models to LLM-based autonomous agents has been driven by advances in tool augmentation, structured planning, and memory architectures, enabling these systems to interact with external knowledge bases and execute clinical information retrieval tasks [1]. In parallel, the adoption of artificial intelligence in medicine has accelerated, with models demonstrating expert-level performance in imaging, pathology, and early warning systems [2], [10]. However, the sensitivity of deep neural networks to adversarial perturbations has been extensively documented. In the textual domain, gradient-based and heuristic search methods can generate input variations that are semantically equivalent to humans but cause consistent misclassification or toxic output generation [3], [6]. The existence of universal adversarial triggers that transfer across models underscores the need for defense mechanisms specifically designed for medical dialogue and documentation contexts, where even a single manipulated recommendation can have grave consequences.

Knowledge graphs offer a complementary paradigm by representing medical entities and their relationships in a structured, queryable format that supports symbolic reasoning and provenance tracking [8]. Previous work has demonstrated the feasibility of constructing health knowledge graphs from electronic medical records, linking symptoms, diseases, medications, and procedures into a coherent ontological framework [4]. Such graphs have been used to augment clinical decision support systems with differential diagnosis suggestions and drug–drug interaction alerts. The integration of knowledge graph retrieval into LLM reasoning has shown promise in reducing hallucination rates and improving factual grounding, yet the vulnerability of graph-augmented architectures to adversarial graph perturbations and misleading relational triples remains under-explored.

On the explainability front, clinicians consistently express a preference for explanations that reference accepted medical knowledge, delineate differential reasoning pathways, and disclose model uncertainty [5]. Techniques such as LIME and SHAP have provided post-hoc feature attribution methods that can highlight which parts of the input influenced a model's output [13], [14]. However, when applied to clinical text generated by LLM agents, these methods often fail to capture the high-level reasoning steps that a physician considers meaningful. Knowledge graph paths, by contrast, naturally align with the diagnostic reasoning processes taught in medical education, offering the possibility of generating chain-of-trust explanations that trace a recommendation back to specific evidence nodes.

Adversarial robustness in medical machine learning has drawn increased attention, with demonstrations that even slight perturbations to medical images or laboratory values can deceive diagnostic classifiers [15]. Extending this concern to LLM agents involves a broader threat surface that includes prompt injection, data poisoning during fine-tuning, model extraction, and adversarial manipulation of retrieved knowledge. Defensive strategies such as adversarial training and certified smoothing have been studied primarily for classification tasks, yet their translation to generative clinical agents remains nontrivial [7], [12]. Researchers have called for defense-in-depth architectures that combine input sanitization, output filtering, and runtime anomaly detection, a perspective that resonates strongly with approaches in cybersecurity for critical infrastructure. Toxicity and safety evaluations of large language models further emphasize the necessity of designing agents not only to resist targeted adversarial inputs but also to maintain robustness under distribution shifts and natural variations in clinical language [9].

3. Architectural Framework for KG-Enhanced Secure LLM Agents

We propose a layered architecture that organizes the clinical LLM agent into five interconnected subsystems: (1) a multimodal input gateway responsible for ingestion and normalization of patient data, clinical notes, and laboratory results; (2) a knowledge graph integration layer that maintains a dynamic, versioned biomedical knowledge graph linking entities from ontologies such as SNOMED CT, RxNorm, and LOINC, enriched with literature-derived relationships; (3) a reasoning core composed of one or more LLM-based agents with tool access, memory, and chain-of-thought scaffolding; (4) a defense-in-depth security fabric that operates across the request, reasoning, and output stages; and (5) an explainability engine that produces structured justification artifacts referencing knowledge graph paths. Each subsystem is designed under the principle that security and explainability are not post-hoc add-ons but intrinsic properties that emerge from the interactions between components.

The input gateway performs normalization and de-identification, but more importantly it hosts a prompt integrity verification module that parses incoming natural language requests to detect adversarial patterns such as obfuscated malicious instructions, repetition-based triggers, and attempts to override system constraints. This module employs a combination of lightweight rule-based filters, perplexity anomaly detection, and a fine-tuned guard model that flags suspicious inputs for human review in high-risk scenarios. Verified inputs are then enriched with structured context by linking clinical concepts to nodes in the knowledge graph, effectively grounding the agent's subsequent reasoning in a curated semantic space before the generation process begins.

The knowledge graph integration layer is the linchpin of the security–explainability synergy. By constraining the agent to reason over a versioned, access-controlled, and continuously validated graph, the architecture reduces the degrees of freedom available to an adversary attempting to steer the agent toward unsafe recommendations. The knowledge graph serves as a structured memory that can be queried through a SPARQL-like interface adapted for LLM function calling, allowing the agent to retrieve clinical guidelines, drug interaction data, and epidemiological trends. Entity linking algorithms resolve ambiguous mentions to canonical graph nodes, mitigating the risk of semantic drift that adversarial inputs often exploit. Moreover, the graph is augmented with provenance metadata, recording the source and update history of each fact, which later feeds into the explainability pipeline.

The reasoning core is composed of a primary LLM agent that integrates retrieved graph subgraphs into its prompt context, using a constrained decoding strategy that penalizes generation of claims unsupported by the retrieved evidence. Multiple specialized agents can be orchestrated for tasks such as differential diagnosis, risk scoring, and treatment planning, with a coordinator agent that resolves conflicts through a voting mechanism informed by knowledge graph consistency scores. The multi-agent coordination introduces redundancy that can be exploited for Byzantine fault tolerance, as long as the knowledge graph itself is not compromised. This design echoes fault-tolerant architectures in distributed systems, where replicas can mask a subset of faulty components.

The defense-in-depth security fabric wraps around the reasoning core with multiple detection and mitigation layers. At the embedding level, adversarial training with perturbed clinical text examples helps the LLM maintain stable representations under small input variations, drawing on techniques similar to those used in visual domain adversarial training [7], [11]. At the semantic level, a runtime monitor compares the logical consistency of the agent's intermediate reasoning steps against the knowledge graph, flagging contradictions as potential

adversarial interference. Finally, an output filter audits recommendations for medical plausibility, drug-allergy conflicts, and compliance with institutional guidelines before they are presented to the clinician. These layers are designed to degrade gracefully; even if an adversary compromises the guard model, the knowledge graph consistency checker provides an independent defense channel.

The explainability engine constructs multi-granular explanations that combine free-text rationales with knowledge graph paths. For a given recommendation, the engine retrieves the subset of graph nodes and edges that were accessed during reasoning and formats them into a human-readable narrative. This approach supports differential explanations, showing why one diagnosis was favored over another by contrasting the supporting evidence paths. Importantly, because the knowledge graph is versioned and auditable, each explanation can be traced to its evidentiary basis, enabling retrospective analysis in the case of adverse outcomes and supporting the continuous certification processes required by evolving medical device regulations.

4. Adversarial Robustness and Defense-in-Depth Strategies

Adversarial threats to clinical LLM agents span multiple layers of the system stack, from the input prompt to the knowledge graph itself. Prompt injection attacks can attempt to override system instructions, forcing the agent to ignore clinical guidelines or disclose protected health information. Semantic perturbation attacks modify the wording of a clinical query in ways that preserve surface meaning for human readers but alter the LLM's internal representations, leading to incorrect triage assessments. More sophisticated adversaries may attempt to poison the knowledge graph by injecting false triples during data ingestion from less-trusted sources. A defense-in-depth posture requires that no single layer is solely responsible for security, and that the system can maintain safety even when some defenses are bypassed.

Recent work has integrated knowledge graph constraints directly into the prompt engineering process, creating a verification loop that rejects adversarial prompts whose inferred intentions conflict with the graph's clinical domain model [16]. This approach leverages the structured semantics of the graph to identify out-of-distribution or maliciously constructed queries, illustrating the defensive value of symbolic oversight. Following this line of reasoning, we extend the concept to a runtime graph-based anomaly detection system that monitors the agent's internal chain-of-thought for deviations from expected reasoning patterns. When the agent proposes a differential diagnosis that contradicts established epidemiological priors encoded in the graph, the system raises an alert and may revert to a safe fallback mode that relies solely on rule-based clinical decision support.

Physical-world adversarial attacks originally demonstrated in computer vision have prompted a reevaluation of robustness for multimodal clinical agents that process images, waveforms, and free text [22]. While a textual adversarial example against an LLM agent may seem simple to generate, the integration of structured lab values and imaging reports introduces coupled perturbation surfaces that require joint defense strategies. Similarly, robustness evaluation benchmarks initially designed for BERT-based classifiers have highlighted that language models can be surprisingly brittle under synonym substitution and character-level perturbations, a finding that generalizes to much larger generative models and demands continuous red-teaming throughout the agent lifecycle [23].

Adversarial training has been a cornerstone of robustness research, yet its application to LLM agents must be carefully calibrated to avoid degrading clinical fluency and factual accuracy.

Training on adversarially crafted clinical notes that contain patient safety-critical perturbations, such as swapped lab values or altered medication names, can harden the model but also risks teaching the agent to be overly skeptical of genuine atypical presentations. The architecture therefore incorporates a dynamic adversarial training curriculum that adapts to the currently deployed threat model, augmented by randomized smoothing at the representation level where feasible. Certified robustness techniques, while still nascent for generative models, offer formal guarantees that can be valuable for specific clinical sub-tasks such as drug interaction classification.

Beyond model-level defenses, the system-level resilience relies on the redundancy introduced by the multi-agent coordination and the knowledge graph integrity protocols. The knowledge graph itself is protected by a combination of digital signatures on fact triples, consensus mechanisms for updates sourced from multiple trusted medical databases, and a versioning system that enables rollback in the event of a detected poison attack. The design philosophy mirrors that of secure distributed ledgers applied to knowledge management, with the crucial difference that the graph must remain rapidly queryable under the latency constraints of clinical workflows. The integration of symbolic verification with neural generation creates a heterogeneous defense surface that raises the cost of successful adversarial compromise by requiring an attacker to simultaneously overcome statistical and logical barriers.

5. Explainability and Trustworthiness in Clinical Decision-Making

Explainability in high-stakes clinical settings is not merely a technical feature but a regulatory requirement and an ethical imperative. Clinicians must be able to interrogate the reasoning behind a recommendation, understand its limits, and contest it when it conflicts with their own clinical judgment. The proposed architecture addresses these needs by generating explanations that are explicitly linked to knowledge graph paths, thereby providing a transparent audit trail. When an agent recommends a specific antibiotic, the explanation engine can surface the relevant guideline node, the pathogen sensitivity data node, and the patient allergy constraints that were considered, presenting a coherent narrative that mirrors differential diagnosis reasoning [5]. This graph-based explanatory paradigm contrasts with pure attention-based heatmaps, which often fail to convey clinically actionable insights.

The trustworthiness of such explanations is intimately tied to the provenance and currency of the underlying knowledge graph. If the graph contains outdated or biased evidence, the explanations will inherit those flaws, potentially reinforcing historical disparities. The architecture therefore mandates a continuous knowledge curation process that incorporates feedback from practicing clinicians, pharmacovigilance databases, and systematic literature reviews. The explainability engine also reports confidence intervals estimated from the agent's ensemble of reasoning paths and flags cases where the knowledge graph contains contradictory evidence, encouraging the clinician to exercise heightened scrutiny. This aligns with the vision of human-AI collaboration in medicine, where the AI serves as a sophisticated second reader rather than an autonomous decision-maker [10].

An important structural trade-off in explainable agent design is between the richness of explanations and the risk of information leakage that could be exploited by adversaries. Detailed explanations that reveal specific reasoning steps and knowledge graph nodes might enable an attacker to craft more effective adversarial prompts by reverse-engineering the agent's internal logic. The architecture manages this tension through a tiered explanation policy: a concise safety-filtered summary is presented by default, while more detailed graph traces are made available to authorized clinicians after authentication and are logged for post-

hoc forensic analysis. This approach respects the dual constraints of clinical transparency and adversarial hardening, acknowledging that the explainability channel itself is a component of the attack surface.

6. Infrastructure, Deployment, and Governance Considerations

Deploying KG-enhanced secure LLM agents within real-world healthcare systems demands careful attention to infrastructure resilience, data governance, and lifecycle management. Hospital information systems are seldom homogeneous; they encompass legacy electronic health records, radiology information systems, laboratory databases, and pharmacy management platforms, each with distinct data schemas and access protocols. The agent's input gateway must be equipped with a robust interoperability layer that employs standardized clinical data models such as the Fast Healthcare Interoperability Resources (FHIR) framework while remaining adaptable to bespoke local configurations. Latency requirements in acute care settings constrain the complexity of graph queries and the number of reasoning hops an agent can perform, necessitating caching strategies for frequently accessed knowledge subgraphs and incremental update mechanisms that avoid full-graph recomputation during clinical hours.

Governance frameworks for AI-based medical devices, including the evolving guidance from regulatory bodies and the European Union AI Act, impose obligations for risk classification, continuous performance monitoring, and human oversight. The proposed architecture addresses these requirements by embedding audit logging at every component interface, from input verification through knowledge graph query results to final recommendations. These logs serve as the evidentiary basis for post-market surveillance and enable the detection of performance drift over time. The versioned knowledge graph facilitates a model-as-medical-device paradigm in which updates can be rolled out incrementally after validation on historical patient cohorts, reducing the regulatory friction associated with full model retraining.

Fairness and equity are critical dimensions of system-level evaluation. Medical knowledge graphs encode not only clinical facts but also the epidemiological patterns present in the training data, which may reflect historical underdiagnosis or overtreatment in specific populations. If left unexamined, graph-enhanced agents could amplify these disparities, as has been documented in risk prediction algorithms that systematically underestimated care needs for Black patients [20]. Mitigating such biases requires a multi-pronged approach: auditing knowledge graph content for representational skew, evaluating agent recommendations across stratified demographic subgroups, and incorporating fairness constraints into the multi-agent coordination mechanism. Research in algorithmic fairness has produced a variety of metrics and debiasing techniques, yet their integration into complex, graph-augmented LLM systems raises unsolved challenges around the trade-off between fairness and accuracy across intersectional groups [21].

Sustainability is another underappreciated factor in the real-world deployment of large-scale clinical AI systems. The computational footprint of continuously running LLM agents, combined with the costs of maintaining and updating a large biomedical knowledge graph, can be substantial. This raises questions about the carbon cost per clinical decision supported and the feasibility of deploying such systems in resource-limited settings. The architecture promotes sustainability through model distillation, sparse expert activation, and edge deployment of lightweight guard models that reduce reliance on centralized cloud resources. Additionally, the knowledge graph's modular design allows facilities to deploy only those

graph modules relevant to their clinical specialty, avoiding unnecessary computational overhead. The long-term viability of secure clinical agents depends on aligning technical sophistication with operational affordability and environmental responsibility.

Policy implications extend beyond the immediate deployment context. The prospect of LLM agents participating in clinical decision-making under adversarial conditions calls for the development of standardized adversarial robustness benchmarks specifically for medical dialogue and documentation tasks. Existing benchmarks address static model accuracy, but they do not adequately capture the dynamic interplay between an agile adversary and a defensive agent system. Regulatory sandboxes and multi-stakeholder consortia that include healthcare providers, technology developers, patient advocacy groups, and cybersecurity experts are needed to establish certification protocols that evolve alongside threat landscapes. Furthermore, international coordination will be essential to harmonize data protection requirements with knowledge sharing across borders, as biomedical knowledge graphs are most valuable when they aggregate diverse global evidence.

7. Conclusion

The integration of knowledge graphs into secure large language model agents represents a promising pathway toward explainable clinical decision support that can withstand adversarial attacks. By weaving together symbolic and subsymbolic reasoning, the proposed architecture addresses the dual challenges of factual grounding and interpretability while embedding defense-in-depth mechanisms that operate at the input, reasoning, and output layers. The system-level analysis reveals that security, explainability, and clinical utility are deeply interdependent properties that cannot be optimized in isolation; strengthening the knowledge graph's provenance tracking, for example, simultaneously enhances forensic audit quality and reduces the attack surface for semantic poisoning. The defense-in-depth strategy, which combines adversarial training, prompt integrity checks, graph-based anomaly detection, and multi-agent redundancy, raises the bar for adversaries and provides graceful degradation pathways.

Looking forward, several areas demand further investigation. The formal verification of combined neural-symbolic pipelines remains in its infancy, and scalable methods for certifying robustness properties of generative clinical agents are urgently needed. The dynamic updating of biomedical knowledge graphs without introducing new vulnerabilities requires secure, community-driven curation protocols akin to those used in open-source software governance. Human–AI interaction studies must move beyond static explanation satisfaction surveys to examine how clinicians detect and respond to adversarial manipulation attempts in real-time clinical workflows. Finally, the global deployment of such systems will require a constellation of policy frameworks that balance innovation with patient safety, algorithmic fairness with clinical accuracy, and data sharing with privacy. The road ahead is demanding, but the convergence of knowledge engineering, adversarial learning, and systems thinking offers a principled foundation for the next generation of trustworthy clinical AI.

References

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

2. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

3. 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 (EMNLP-IJCNLP) (pp. 2153–2162).

4. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. Scientific Reports, 7(1), 5994.

5. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In Machine Learning for Healthcare Conference (pp. 359–380). PMLR.

6. Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2021–2031).

7. 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.

8. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11–33.

9. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3356–3369).

10. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

11. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

12. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (pp. 1310–1320).

13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

15. 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.

16. 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.

17. Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web (pp. 697–706).

18. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (pp. 1247–1250).

19. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., ... & Zimmermann, A. (2022). Knowledge graphs. ACM Computing Surveys, 54(4), 1–37.

20. 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.

21. 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.

22. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1625–1634).

23. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 8018–8025).

24. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

25. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094.