Graph-Enhanced Clinical Knowledge Injection for Secure and Robust Medical LLM Reasoning

Tobias D. Greene
Department of Computer Science, George Mason University, Fairfax, VA, USA.
tobiasgreene@gmu.edu

Zixuanan Wan
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
wanzixuanan@colostate.edu

Ananya M. Chopra
Department of Computer Science, University of North Texas, Denton, TX, USA.
ananya120@unt.edu

Kiran Chatterjee
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
kiranc@oregonstate.edu

Abstract

Large language models have demonstrated remarkable capabilities in natural language understanding and generation, yet their deployment in high-stakes medical contexts remains fraught with challenges related to factual accuracy, adversarial vulnerability, and systemic bias. This paper proposes a graph-enhanced clinical knowledge injection framework that systematically integrates structured biomedical ontologies and relational knowledge graphs into the reasoning pipeline of medical large language models. By embedding graph-based representations of clinical entities, relationships, and hierarchical dependencies, the proposed architecture augments model outputs with verifiable domain constraints and causal pathways, thereby improving both security and robustness. We examine the architectural trade-offs between expressivity and computational overhead, the role of graph neural network layers in preserving semantic integrity, and the implications for adversarial robustness when knowledge graphs serve as external verifiers. The framework is situated within a broader governance perspective that addresses data provenance, fairness across demographic groups, and sustainability of large-scale inference. Through cross-domain comparisons with existing retrieval-augmented generation and fine-tuning approaches, we highlight the structural advantages of graph-enhanced injection for mitigating hallucinations and resisting malicious perturbations. The paper concludes with a forward-looking discussion on the deployment of such systems in clinical decision support, the need for continuous validation against evolving medical knowledge, and the policy infrastructure required to ensure equitable access and accountability.

Keywords

graph-enhanced reasoning, clinical knowledge injection, medical large language models, adversarial robustness, knowledge graphs, secure AI, healthcare AI governance.

1. Introduction

The integration of large language models into clinical workflows promises to transform medical diagnosis, treatment planning, and patient communication. However, the raw statistical learning paradigm underlying these models introduces fundamental risks: hallucinations, sensitivity to input perturbations, and a lack of explicit reasoning over structured clinical knowledge. These vulnerabilities are particularly acute in medicine, where erroneous outputs can lead to patient harm and erode trust in automated systems [1, 2]. Addressing these challenges requires not only better training data and alignment techniques but also architectural interventions that embed domain-specific knowledge in a verifiable and updatable manner.

Graph-based representations offer a natural medium for encoding the relational and hierarchical nature of clinical knowledge. Medical ontologies such as SNOMED CT, ICD-10, and the Gene Ontology define thousands of entities with complex interconnections that reflect causal, taxonomic, and procedural relationships [3, 4]. By injecting such graph-structured knowledge into large language models, it becomes possible to constrain the model's reasoning space, enhance factual consistency, and provide a mechanism for post-hoc verification. This approach differs from retrieval-augmented generation, which typically retrieves text snippets, and from fine-tuning, which modifies model weights without explicit structural guidance [5, 6].

In this paper, we propose a graph-enhanced clinical knowledge injection framework that combines a pretrained language model with a graph neural network encoder and a knowledge graph database. The graph encoder processes structured clinical relationships and generates embeddings that are fused with the language model's hidden states during inference. This fusion enables the model to attend to both textual context and formal knowledge, thereby improving robustness against adversarial inputs that attempt to exploit missing or contradictory information. We systematically analyze the architectural design choices, the security benefits of graph-based verification, and the broader socio-technical implications for safe deployment in healthcare settings.

2. Background and Motivation

Medical decision-making tasks require reasoning over a vast and evolving body of knowledge that is inherently structured: diseases have symptoms, drugs have contraindications, and procedures have indications. Large language models trained predominantly on unstructured text often fail to capture these relational dependencies, leading to errors that range from minor inaccuracies to clinically dangerous misdiagnoses [7]. Recent studies have documented that even state-of-the-art models exhibit high rates of hallucination when asked about rare conditions or drug interactions [8]. These failures underscore the need for mechanisms that ground model outputs in authoritative knowledge sources.

Retrieval-augmented generation has emerged as a popular solution, where relevant documents are fetched from a knowledge base and appended to the input context [9]. While effective for factoid questions, retrieval-augmented generation struggles with multi-hop reasoning that requires traversing relational paths, such as inferring a drug's side effect through an intermediate condition. Moreover, retrieval-augmented generation is vulnerable to adversarial retrieval, where malicious queries cause the retriever to select misleading or poisoned documents [10]. Fine-tuning on domain-specific corpora can improve coverage but does not guarantee structural consistency and may introduce catastrophic forgetting of general knowledge [11].

Graph-based knowledge injection offers a complementary paradigm. By representing clinical knowledge as a graph where nodes correspond to entities and edges to relationships, the model can perform reasoning that is both explicit and compositional [12]. Early work on graph neural network-augmented language models has shown improvements in tasks such as relation extraction and question answering on biomedical benchmarks [13]. However, the specific application to security and robustness in medical reasoning remains underexplored. As adversarial threats become more sophisticated, including prompt injection attacks and data poisoning, the ability to cross-reference outputs against a trusted knowledge graph becomes a critical defensive layer [14].

The proposed framework in this paper builds on these foundations by integrating graph knowledge injection directly into the inference pipeline of a medical large language model. This integration not only enhances factual accuracy but also provides a structural basis for detecting and mitigating adversarial manipulations. The following sections detail the architecture, its security properties, and the governance considerations that accompany its real-world deployment.

3. Graph-Enhanced Knowledge Injection Architecture

The core architecture consists of three interconnected modules: a pretrained language model backbone, a graph neural network encoder, and a fused reasoning layer. The language model, which may be a decoder-only Transformer such as GPT-4 or LLaMA, handles the textual input and generates intermediate representations. In parallel, a clinical knowledge graph containing entities and relationships from standardized ontologies is encoded using a graph neural network that produces vector embeddings for each node and edge. These graph embeddings are then injected into the language model's hidden states at selected transformer layers via a gating mechanism that learns to balance the contribution of textual and graph-based signals.

The choice of graph neural network architecture is critical for preserving the relational semantics of medical knowledge. Graph attention networks, which assign adaptive weights to neighboring nodes, are particularly suitable because they can model the varying importance of different clinical relationships, such as the strong causal link between a pathogen and a disease versus a weaker associative link between a symptom and a comorbidity [15]. Additionally, relational graph convolutional networks can capture edge types explicitly, allowing the model to distinguish between hierarchical relations like "is-a" and procedural relations like "treats" [16]. The graph neural network is trained jointly with the language model using a multi-task objective that combines standard language modeling loss with a graph reconstruction loss, ensuring that the graph embeddings remain faithful to the ontology structure.

During inference, the fused representations are passed to a reasoning head that generates the final output. An important design consideration is the latency and memory overhead introduced by graph encoding. Clinical knowledge graphs can be extremely large, with SNOMED CT alone containing over 350,000 concepts. To manage computational cost, the framework employs a subgraph sampling strategy that extracts only the most relevant entities and relationships based on the input context [17]. This is achieved through a lightweight entity linking module that identifies mentions in the user query and retrieves the corresponding neighborhoods from the knowledge graph. The sampled subgraph is then encoded on-the-fly, enabling real-time inference without requiring the entire graph to be loaded into memory.

Another architectural trade-off involves the depth of graph injection. Injecting graph embeddings at early transformer layers allows the model to condition its entire generation on structured knowledge, but risks over-constraining creativity for open-ended questions. Injecting at later layers, closer to the output, provides a final verification step without altering the core language understanding. Empirical studies on biomedical question answering suggest that a hybrid approach, where graph information is injected at both early and late layers with different weighting, yields the best balance between accuracy and flexibility [18]. The framework incorporates learnable gates that automatically adjust these weights based on the complexity of the input, offering a form of dynamic architecture.

4. Security and Robustness Considerations

The primary security benefit of graph-enhanced knowledge injection lies in its ability to provide an external, verifiable reference for model outputs. Adversarial attacks on large language models often exploit the model's reliance on statistical patterns rather than true understanding. For example, a prompt injection attack might embed a malicious instruction that causes the model to ignore safety guidelines and generate harmful medical advice. In the proposed framework, the graph encoder provides a constraint that cannot be easily bypassed by textual perturbations because the graph embeddings are derived from a static, curated ontology [14]. Even if the language model's textual representations are corrupted, the fused output must still be consistent with the graph's relational structure. This creates a form of reasoning redundancy that increases the difficulty of successful attacks.

Another class of adversarial threats involves data poisoning, where training data is contaminated with incorrect medical facts. Because the graph knowledge is injected after pretraining and is not derived from the same corpus, it acts as an independent source of truth. If the language model attempts to recall a poisoned fact that contradicts the graph, the fusion mechanism can downweight that output or flag an inconsistency for human review. This property is especially valuable in medical settings where the cost of false positives and false negatives is asymmetric. The framework can be configured with a confidence threshold: when the graph-based and text-based predictions diverge significantly, the system abstains from answering and defers to a clinician [19].

Robustness to distributional shift is another advantage. Medical knowledge evolves over time, and models trained on static datasets become outdated. In the graph-enhanced framework, updating the knowledge graph is sufficient to refresh the model's reasoning, without requiring retraining of the entire language model. This modularity reduces the carbon footprint and computational cost of continuous deployment, aligning with sustainability goals [20]. Moreover, the graph can be versioned and audited, providing a transparent record of which clinical relationships were available at the time of inference. This audit trail is essential for medico-legal accountability.

However, the graph-enhanced approach is not immune to adversarial manipulation of the knowledge graph itself. An attacker with write access to the graph database could introduce spurious edges or inject malicious nodes that mislead the reasoning process. Defending against such attacks requires robust graph maintenance protocols, including cryptographic signatures for provenance, regular consensus checks against authoritative sources, and anomaly detection on graph updates [21]. The framework should also incorporate adversarial training on simulated graph perturbations to make the fusion layer resilient to minor inaccuracies. These considerations highlight that security is a system-level property that extends beyond the algorithm to the data infrastructure and organizational governance.

5. Governance and Deployment Implications

Deploying a graph-enhanced medical large language model in clinical settings raises numerous governance challenges that must be addressed before widespread adoption. The first pertains to data provenance and intellectual property. Clinical knowledge graphs are often derived from proprietary ontologies developed by medical associations or commercial vendors. Licensing terms may restrict redistribution or modification, complicating the deployment of open-source variants [22]. Healthcare institutions must negotiate clear usage rights and establish data-sharing agreements that respect patient privacy and comply with regulations such as HIPAA and GDPR.

Fairness is another critical dimension. Medical knowledge graphs may encode biases present in the underlying biomedical literature, which historically underrepresents certain populations [23]. For example, disease prevalence data derived from predominantly Caucasian cohorts can lead to graph edges that misrepresent symptom-disease associations for other racial groups. The graph injection mechanism amplifies these biases if left unchecked. Mitigation strategies include diversifying the ontology sources, incorporating community health data, and applying fairness-aware graph neural network training that penalizes disparities in prediction accuracy across demographic subgroups [24]. The framework should include a fairness auditing module that periodically evaluates the sensitivity of outputs to demographic variables.

Sustainability and energy efficiency are increasingly important for large-scale AI systems. The graph-enhanced architecture introduces additional computational overhead for graph encoding and subgraph sampling. To reduce environmental impact, the framework can leverage sparse graph representations and quantization techniques that lower memory and energy consumption without sacrificing accuracy [25]. Furthermore, the modular design allows the language model backbone to be a smaller, distilled model because the graph compensates for missing knowledge. This trade-off between model size and graph complexity should be evaluated on a per-application basis to minimize total carbon footprint.

Finally, the deployment of such a system requires a human-in-the-loop governance structure. Clinical decision support tools must not replace physician judgment but augment it. The graph-enhanced framework can be designed to output not only a recommendation but also a traceable reasoning path that shows which graph edges were used to arrive at the conclusion. This transparency enables clinicians to verify the logic and override the system when necessary. Policies should mandate that all AI-generated suggestions be accompanied by confidence scores and citations to the underlying knowledge graph nodes, facilitating second-opinion workflows. Regulatory bodies, such as the FDA for clinical decision support software, may need to establish new standards for graph-based AI systems that account for the dynamic nature of knowledge graphs and the need for continuous validation.

6. Conclusion

This paper has presented a graph-enhanced clinical knowledge injection framework designed to improve the security and robustness of medical large language model reasoning. By embedding structured ontological knowledge into the inference pipeline via graph neural network encoders, the framework provides a verifiable constraint that mitigates hallucinations, resists adversarial attacks, and supports continual updates without full model retraining. The architectural discussion highlighted trade-offs between expressivity, latency, and computational cost, while the security analysis underscored the importance of independent

knowledge sources for defense against prompt injection and data poisoning. Governance considerations related to provenance, fairness, sustainability, and human oversight have been addressed to frame the technology within a realistic deployment context.

Future research should explore the integration of dynamic knowledge graphs that incorporate real-time clinical trial results and patient outcome data, as well as federated learning approaches that allow multiple institutions to contribute graph updates without sharing sensitive patient information. The interplay between graph-based reasoning and emerging techniques such as chain-of-thought prompting and tool-use remains a fertile area for investigation. As medical AI systems transition from research prototypes to clinical tools, the principles of structural knowledge injection advocated here will be essential for building trust, ensuring accountability, and ultimately improving patient outcomes.

References

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

2. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940.

3. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(suppl_1), D267-D270.

4. Spackman, K. A., Campbell, K. E., & Cote, R. A. (1997). SNOMED RT: a reference terminology for health care. Journal of the American Medical Informatics Association, 4(Suppl), 640-644.

5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

6. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 328-339.

7. Wang, B., Xie, Q., Pei, J., Lee, M. T., & Chen, C. (2023). Are large language models ready for healthcare? A comprehensive review. arXiv preprint arXiv:2304.09685.

8. Asakura, K., Kaneko, M., & Aizawa, A. (2024). Hallucination detection in medical LLMs: a survey. Journal of Biomedical Informatics, 152, 104621.

9. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: retrieval-augmented language model pre-training. Proceedings of the 37th International Conference on Machine Learning, 3929-3938.

10. Chen, H., Ji, H., & Roth, D. (2021). Adversarial retrieval for open-domain question answering. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 4210-4221.

11. McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. Psychology of Learning and Motivation, 24, 109-165.

12. Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2020). BioBERT: a pre-trained biomedical language model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

13. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., & Sun, M. (2020). Graph neural networks: a review of methods and applications. AI Open, 1, 57-81.

14. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

15. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations.

16. Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. European Semantic Web Conference, 593-607.

17. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30.

18. Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2022). QA-GNN: reasoning with language models and knowledge graphs for question answering. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, 535-547.

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

20. 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, 3645-3650.

21. Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems, 30.

22. Smith, B., & Ceusters, W. (2010). Ontologies and the semantic web. Applied Ontology, 5(3-4), 155-175.

23. Cahan, E. M., & Hernandez-Boussard, T. (2023). Bias in machine learning for health: a review of the literature and recommendations for future work. JAMA Health Forum, 4(5), e231087.

24. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. Proceedings of the 35th International Conference on Machine Learning, 60-69.

25. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.