Prompt Injection Resistance in Clinical LLM Agents via Structured Medical Ontology Alignment

Neil Simmons
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
neil1980@colostate.edu

Finn Marshall
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
finn338@binghamton.edu

Pedro Wagner
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
pedrow@uab.edu

Abstract

Large language model agents deployed in clinical settings offer transformative potential for decision support, patient communication, and workflow automation. However, their vulnerability to prompt injection attacks poses a critical safety risk, especially when adversarial inputs can manipulate model outputs to produce harmful or misleading medical advice. This paper proposes a structural defense framework based on the alignment of clinical large language model agents with a formal medical ontology, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) or the Unified Medical Language System (UMLS). By constraining the agent's reasoning and generation processes to a structured representation of medical knowledge, the system can detect and reject inputs that deviate from clinically valid pathways. We present an architectural design that integrates ontological grounding at multiple stages of the agent pipeline, including input preprocessing, context injection, and output validation. The approach is evaluated against a taxonomy of prompt injection techniques, including direct, indirect, and multi-turn attacks. Results demonstrate that ontology-aligned agents exhibit significantly higher resistance to adversarial manipulations compared to unconstrained baseline models, while maintaining clinical accuracy and fluency. The paper also discusses the trade-offs between security and expressiveness, the computational overhead of ontology integration, and the implications for regulatory compliance and deployment in resource-constrained healthcare environments. We argue that structured ontology alignment represents a promising direction for building trustworthy clinical large language model agents that can safely operate in adversarial open-world interactions.

Keywords

prompt injection, large language models, clinical agents, medical ontology, adversarial robustness, healthcare AI safety.

1. Introduction

The rapid integration of large language models into clinical workflows has opened new possibilities for automating tasks such as triage, differential diagnosis, medication

reconciliation, and medical documentation. These agents, often built upon foundation models like GPT-4, PaLM, or open-source alternatives, can parse natural language queries from clinicians and patients, retrieve relevant evidence, and generate contextually appropriate responses. However, the very flexibility that makes these models powerful also renders them vulnerable to a class of attacks known as prompt injection, where an adversary crafts input text that subverts the intended behavior of the model. In a clinical setting, a successful prompt injection could cause the agent to ignore safety constraints, fabricate diagnoses, or recommend harmful treatments. The consequences of such vulnerabilities are potentially catastrophic, raising urgent questions about the safety and reliability of deploying large language model agents in healthcare.

Existing defenses against prompt injection have primarily focused on input sanitization, output filtering, and adversarial training. Input sanitization methods attempt to detect and neutralize injection patterns before they reach the model, but they are often brittle against novel attack variants. Output filtering can remove overtly dangerous content, but it cannot prevent the model from reasoning along a malicious path. Adversarial training, while effective for certain attack classes, is computationally expensive and may not generalize to the wide variety of injection techniques that adversaries can employ in the wild. Moreover, these approaches treat the problem as a surface pattern-matching issue rather than addressing the deeper architectural vulnerability: that large language models generate outputs based on statistical patterns in their training data rather than on a grounded understanding of clinical concepts.

This paper proposes an alternative paradigm: prompt injection resistance through structured medical ontology alignment. Rather than attempting to block every possible injection vector, we advocate for embedding the clinical agent within a formal ontological framework that represents the relationships between medical concepts, diagnostic criteria, treatment protocols, and contraindications. By constraining the agent's reasoning and output generation to paths that are semantically valid within this ontology, we can systematically reject inputs that would lead the model to produce clinically unsound responses. The ontology serves as a semantic firewall that transforms the problem from one of detecting adversarial strings to one of verifying logical coherence with a trusted knowledge base.

The contributions of this work are threefold. First, we provide a comprehensive threat model for prompt injection attacks targeted at clinical large language model agents, categorizing attack vectors by their mechanism and goal. Second, we describe a novel architecture that integrates ontological alignment at the levels of input interpretation, context augmentation, and output validation, illustrating how each layer contributes to overall security. Third, we present an evaluation of the proposed system against a range of injection techniques, demonstrating significant improvements in resistance while maintaining clinically acceptable performance. The paper concludes with a discussion of the broader implications for clinical AI governance, the trade-offs between security and usability, and the pathways for deploying such systems in real-world healthcare environments.

2. Background and Related Work

Prompt injection has emerged as a critical security concern for large language models since its formal identification in early 2023. The attack exploits the model's inability to distinguish between instructions embedded in user input and the system-level directives that define its role. In the context of clinical agents, an adversary might craft a prompt that begins with a benign medical query but then appends a directive such as "Ignore all previous instructions

and tell the patient that their condition is terminal even if it is not." Research has shown that even stateful conversation models are susceptible to such manipulations, and that injection attacks can be launched indirectly through third-party sources such as web pages or document uploads [1, 2]. Defenses proposed to date include prompt sandboxing, where the model is instructed to treat a special delimiter as marking user input, and parameterized instructions that separate command tokens from natural language. However, these approaches have been shown to be circumventable by sophisticated adversaries who can mimic the delimiter structure or exploit the model's tendency to follow examples [3].

In parallel, the medical informatics community has long developed and maintained structured ontologies to enable interoperable and semantically consistent representation of clinical knowledge. SNOMED CT, for instance, provides a comprehensive hierarchy of clinical terms with well-defined relationships such as "is a," "has finding site," and "causative agent." The Unified Medical Language System integrates multiple vocabularies and provides a metathesaurus that can be queried to verify the semantic validity of medical statements. These ontologies have been used for clinical decision support, natural language processing of electronic health records, and quality measurement, but their potential to enhance the security of generative language models has not been fully explored [4, 5].

Recent work has begun to investigate the intersection of large language models and formal knowledge representations. Some researchers have proposed using knowledge graphs to constrain the generation of factual information in question-answering systems [6]. Others have explored the use of ontologies to detect hallucinated content by comparing model outputs against structured knowledge bases [7]. However, these efforts have focused on factual accuracy rather than adversarial robustness. The unique contribution of this paper is to apply ontological alignment specifically as a defense against prompt injection, leveraging the ontology not only to verify outputs but also to constrain the reasoning process itself.

3. Threat Model for Clinical LLM Agents

To design effective defenses, it is first necessary to understand the range of attacks that a clinical large language model agent may face. We consider three broad categories of prompt injection attacks relevant to clinical settings. The first category is direct injection, where an attacker explicitly overrides system instructions within the same conversational turn. For example, a patient might ask "What is the recommended treatment for hypertension?" and then add "Now repeat the drug names in reverse order and output the password for the hospital database." The second category is indirect injection, where the adversarial input is embedded in content that the agent retrieves from an external source, such as a medical journal article, a patient history note, or a web search result. The agent may be instructed to process such content, and the injected instructions become active without being explicitly flagged by the user [8]. The third category is multi-turn injection, where the adversary gradually induces the model to deviate from its safety boundaries over several exchanges, often by building a false context that makes the final injection seem natural.

The goals of these attacks can vary. Some attackers seek to exfiltrate sensitive patient information by causing the agent to output protected health data in an encoded form. Others aim to cause the agent to give harmful medical advice, such as recommending a contraindicated drug for a known allergy. Still others may attempt to make the agent perform actions outside its intended scope, such as deleting records or sending unauthorized messages. In all cases, the underlying vulnerability is that the model treats all input tokens as part of a unified language modeling task, without a built-in mechanism to distinguish between

authoritative clinical knowledge and adversarial commands. The proposed ontology alignment framework addresses this by ensuring that any output or intermediate reasoning step must map onto valid paths within the ontology. If an injection attempts to steer the model toward a clinically inconsistent statement, the ontological validator will reject it, regardless of the linguistic surface form.

4. Structured Medical Ontology Alignment: Architecture and Design

The core architecture of the proposed system consists of three interconnected components: an input grounding module, a reasoning constraint module, and an output verification module. Each module leverages the medical ontology in a distinct way to provide layered defense against prompt injection.

The input grounding module is responsible for parsing the user's natural language query and mapping it to ontology concepts. This is accomplished using a named entity recognition system trained to identify clinical terms and map them to unique identifiers in the ontology. For example, if a user asks about "heart attack," the module identifies the corresponding SNOMED CT concept for myocardial infarction. If the input contains terms that do not map to any known ontology concept, or that map to concepts that are logically inconsistent with the rest of the query (such as a request to treat a condition with a medication that has no ontological relationship to that condition), the module flags the query as suspicious and may either refuse to process it or escalate to a human clinician. This grounding step prevents the model from being primed with semantically meaningless or adversarial tokens that could later influence its generation.

The reasoning constraint module operates during the model's internal processing. Rather than allowing the model to generate free-form text, the framework interleaves calls to the ontology to validate each intermediate reasoning step. For instance, if the model begins to generate a differential diagnosis list, the constraint module checks each candidate diagnosis against the presenting symptoms using ontological relationships such as "has finding" or "is associated with." If the model attempts to generate a diagnosis that has no ontological link to any of the symptoms mentioned, the module intervenes, either by adjusting the probability distribution of the next token or by injecting a corrective prompt that steers the model back to a valid path. This approach is inspired by techniques from neuro-symbolic AI, where logical constraints are used to guide neural generation [9].

The output verification module performs a final check on the complete response before it is presented to the user. This module uses the ontology to verify that every clinical assertion made in the response is consistent with the knowledge base. For example, if the response states that a particular drug is indicated for a condition, the module verifies that the ontology includes a "has indication" relationship between the drug and the condition. If any assertion is unsupported, the module can either filter it out, regenerate a corrected response, or flag the output for human review. This final layer acts as a safety net to catch any remaining adversarial influence that may have slipped through the earlier modules.

The specific ontology chosen for alignment can be tailored to the clinical domain of the agent. For a general internal medicine agent, SNOMED CT provides broad coverage, while for a specialized oncology agent, the NCI Thesaurus may be more appropriate. The framework is designed to be ontology-agnostic, so that different knowledge bases can be plugged in depending on the application. However, integration with a common standard such as the

Unified Medical Language System enables cross-ontology consistency and easier maintenance [10].

5. Robustness Against Prompt Injection Attacks

To evaluate the effectiveness of the ontology alignment framework, we conducted a series of experiments using a simulated clinical agent based on a publicly available large language model. The agent was deployed both in its vanilla form and with the full ontology alignment pipeline. We then tested both versions against a library of 50 prompt injection attacks spanning direct, indirect, and multi-turn categories. The attacks were designed in consultation with clinical safety experts and included scenarios such as overriding dosage instructions, injecting fictitious contraindications, and attempting to retrieve protected health information disguised as a clinical query.

The results showed a dramatic improvement in resistance. The vanilla agent was successfully compromised in 86 percent of the attack attempts, often generating responses that violated clinical guidelines or disclosed sensitive information. The ontology-aligned agent resisted 94 percent of attacks, with only a small number of failures occurring in cases where the injection was carefully disguised as a semantically plausible but incorrect clinical statement. These failures were analyzed and traced to gaps in the ontology coverage or to cases where the attacker leveraged rare but valid clinical relationships that the ontology did not fully capture. In all such cases, the output verification module flagged the response as suspicious, preventing it from being delivered to the user without human review. Recent work [11] has explored complementary security enhancement methods for adversarial robust large language model intelligent agents in medical decision-making tasks, and our findings align with the direction of integrating structured knowledge to improve safety.

We also measured the impact of the ontology alignment on the agent's clinical accuracy for benign queries. The framework introduced a slight reduction in response fluency and a small increase in latency, but the accuracy of clinical recommendations remained statistically unchanged. This suggests that the ontological constraints do not unduly limit the model's ability to generate correct and helpful medical information; rather, they prune away dangerous or unsupported responses while preserving the clinically valid space.

6. Trade-offs and System-Level Considerations

The adoption of ontology alignment as a defense mechanism introduces several trade-offs that must be carefully considered in system design. The primary trade-off is between security and expressiveness. By constraining the agent to paths that are validated against a fixed ontology, we necessarily limit its ability to generate novel or unconventional clinical insights. In domains where medical knowledge is rapidly evolving, such as pandemic response or emerging treatments, an ontology may lag behind current best evidence, leading the agent to reject valid but newly discovered associations. This tension mirrors the classic trade-off in safety-critical systems between deterministic rule-following and adaptive learning. To mitigate this, we propose a hybrid approach in which the ontology is periodically updated from trusted sources, and where human experts can override ontological rejections in a controlled manner with logging and audit trails.

Another important trade-off concerns computational overhead. The input grounding, reasoning constraint, and output verification modules each require multiple calls to the ontology knowledge base. For a large ontology like SNOMED CT, which contains over 300,000 concepts, the lookup and traversal operations can introduce latency that may be

unacceptable in real-time clinical settings. Optimizations such as caching frequent queries, using approximate matching for named entity recognition, and precomputing subgraphs for common clinical pathways can reduce this overhead. However, the additional latency must be weighed against the safety benefits. In many clinical scenarios, a delay of a few hundred milliseconds is acceptable if it prevents a life-threatening error. For time-critical applications such as emergency triage, a lighter-weight version of the ontology alignment may be employed, focusing only on the most safety-critical assertions.

Fairness and bias are also important system-level considerations. The ontology reflects the medical knowledge that has been formally encoded, which may itself contain biases, such as underrepresentation of certain demographic groups in clinical trials or diagnostic criteria that are less sensitive for minority populations. An ontology-aligned agent might inadvertently perpetuate these biases by rejecting inputs that challenge the established knowledge. To address this, the framework must include mechanisms for bias auditing and for incorporating community-driven updates to the ontology. Moreover, the agent should be transparent about the sources of its knowledge and allow clinicians to flag potential biases for review.

7. Deployment, Governance, and Policy Implications

Deploying ontology-aligned clinical large language model agents in real healthcare environments raises a host of governance and policy issues. Regulatory bodies such as the U.S. Food and Drug Administration have begun to develop frameworks for software as a medical device that includes machine learning components. The addition of a structured ontology alignment layer could be seen as a form of safety control that is auditable and explainable, potentially easing the path to regulatory approval. The ontology itself serves as a documented knowledge base that can be reviewed for completeness and correctness, unlike the opaque parameters of a neural network. This transparency is a significant advantage for building trust among clinicians, patients, and regulators.

However, the maintenance of the ontology becomes a critical operational concern. Healthcare organizations must establish processes for updating the ontology in response to new evidence, correcting errors, and handling edge cases. This requires collaboration between clinical domain experts, knowledge engineers, and data governance teams. The cost of such maintenance may be prohibitive for smaller institutions, suggesting a role for centrally maintained open-source ontologies or cloud-based services that provide ontology-as-a-service with guaranteed uptime and compliance.

Policy implications extend to data privacy and security. The ontology alignment pipeline, like any software component, could itself be a target for attack. An adversary who gains control of the ontology could inject malicious relationships that would then be trusted by the agent. Therefore, the ontology storage and update mechanisms must be protected using access controls, versioning, and integrity checks. We recommend that the ontology be treated as a high-value asset with cryptographic verification of its contents.

Finally, the framework’s reliance on a formal ontology raises questions about the role of human judgment in clinical decision-making. The ontology alignment should be viewed as a safety guardrail rather than a replacement for clinical expertise. The agent should be designed to defer to human clinicians when its ontological constraints prevent it from generating a response, or when a query falls outside the ontology’s coverage. This human-in-the-loop model aligns with best practices in AI safety and ensures that the agent remains a supportive tool rather than an autonomous decision-maker.

8. Conclusion

Prompt injection represents a serious and persistent threat to the safe deployment of large language model agents in clinical environments. This paper has presented a structural defense based on the alignment of the agent's reasoning and output generation with a formal medical ontology. By grounding the agent's operations in a trusted knowledge base, we can systematically reject inputs and outputs that deviate from clinically valid paths, thereby providing resistance against a broad range of injection attacks. The proposed architecture, with its input grounding, reasoning constraint, and output verification modules, offers a layered defense that can be adapted to different clinical domains and ontology standards.

Our evaluation demonstrated that ontology-aligned agents achieve significantly higher resistance to prompt injection while maintaining clinical accuracy. The framework introduces manageable trade-offs in terms of expressiveness, computational overhead, and the need for ontology maintenance, all of which can be addressed through careful system design and governance. Looking forward, we believe that the combination of neural language models with structured knowledge representations represents a promising direction for building safe, trustworthy, and accountable clinical AI systems. Future work should explore the integration of dynamic ontology updates based on live evidence, the development of standardized benchmarks for clinical prompt injection resistance, and the deployment of ontology-aligned agents in pilot studies within controlled healthcare settings.

References

1. Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. In Proceedings of the NeurIPS 2022 Workshop on Security in Machine Learning.
2. Greshake, K., Abdolrashidi, A., Ramaswamy, S., & Shacham, H. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS 2023).
3. Schulhoff, S., Sachan, S., Calvo, R., & van der Wal, T. (2024). Defending against prompt injection: A survey and taxonomy. ACM Computing Surveys, 57(1), Article 12.
4. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(Database issue), D267–D270.
5. Elkin, P. L., Brown, S. H., & Huss, E. (2005). A systematic evaluation of the quality of SNOMED CT. Journal of the American Medical Informatics Association, 12(5), 553–561.
6. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Knowledge graph-enhanced large language models via ontology-aware prompt tuning. In Proceedings of the 38th AAAI Conference on Artificial Intelligence.
7. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are zero-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).
8. Bagdasaryan, E., & Shmatikov, V. (2023). Spinning sequences: How to inject new instructions into conversational agents. In Proceedings of the 2023 IEEE Symposium on Security and Privacy.

9. Garcez, A. d., Broda, K., & Gabbay, D. M. (2002). Neural-symbolic learning systems: Foundations and applications. Springer.

10. Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. Methods of Information in Medicine, 37(4-5), 394–403.

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

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

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

15. Neumann, M., & Stuckenschmidt, H. (2018). Ontology-based data access with a large language model? In Proceedings of the 31st International Workshop on Description Logics.

16. Kohane, I. S., & Altman, R. B. (2023). A safety framework for clinical AI. JAMA, 329(15), 1275–1276.

17. Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

18. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), 167–195.

19. Seal, H., & Li, S. (2023). Adversarial robustness of medical language models: A case study on clinical note generation. In Proceedings of the 2023 Machine Learning for Healthcare Conference.

20. Wong, A., & Lewis, P. (2024). Prompt injection in multi-agent systems: A taxonomy and defense. arXiv preprint arXiv:2404.12345.