PathGuard-Med: Interpretable Safety Alignment for Clinical Large Language Models via Multi-Hop Reasoning Intervention

Jakub L. Rao
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
jlrao@unh.edu

Abstract

The deployment of large language models in clinical settings promises transformative improvements in diagnostic support, patient communication, and administrative efficiency, yet it simultaneously introduces profound risks related to patient safety, interpretability, and regulatory compliance. Existing safety alignment techniques, such as reinforcement learning from human feedback and constitutional AI, primarily operate at the output level, penalizing harmful responses without providing transparent mechanisms for why a given response was deemed unsafe or how it could be corrected. This paper introduces PathGuard-Med, a novel interpretable safety alignment framework designed specifically for clinical large language models. PathGuard-Med leverages multi-hop reasoning intervention to trace and adjust the internal reasoning pathways of a model before a response is generated, thereby enabling clinicians and compliance officers to inspect, validate, and override safety decisions in real time. The framework integrates a multi-hop graph structure that encodes clinical knowledge, ethical constraints, and regulatory guidelines into interleaved reasoning chains. When a query enters the system, PathGuard-Med routes the computation through a series of intervention points where explicit reasoning steps are monitored and, if necessary, redirected to avoid unsafe conclusions. This architecture does not merely filter outputs but restructures the model's reasoning process, making safety alignment both interpretable and auditable. We discuss the structural trade-offs between intervention granularity and computational overhead, the governance challenges of deploying such a system in hospital networks, and the policy implications for algorithmic accountability under frameworks like the FDA's Software as a Medical Device guidelines. Through a comparative analysis with existing approaches, we demonstrate that PathGuard-Med achieves higher transparency without sacrificing clinical utility. The paper concludes with forward-looking perspectives on embedding interpretable safety mechanisms into next-generation clinical AI infrastructures.

Keywords

clinical large language models, interpretable safety alignment, multi-hop reasoning, intervention architecture, healthcare AI governance.

1. Introduction

The rapid integration of large language models into clinical workflows has generated both enthusiasm and caution among healthcare stakeholders. These models, when fine-tuned on medical corpora, can assist in summarizing patient histories, generating differential diagnoses, and even drafting clinical notes [1]. However, the stakes in clinical environments are uniquely high: an incorrect or unsafe recommendation can directly lead to patient harm, legal liability, and erosion of trust in AI-assisted medicine [2]. Traditional safety alignment methods, such as

reinforcement learning from human feedback (RLHF) and constitutional AI, have demonstrated effectiveness in reducing overtly harmful outputs in general-purpose chatbots, but they suffer from several critical limitations when transferred to clinical contexts [3]. RLHF, for instance, relies on human annotators scoring model outputs, which introduces subjectivity and scalability issues, while constitutional AI encodes rules as static principles that cannot adapt to the nuanced ethical dilemmas present in clinical decision-making [4]. More fundamentally, these methods operate post-hoc, intervening only after the model has already generated an output, thereby obscuring the internal reasoning that led to the unsafe response. This opacity is particularly problematic in medicine, where clinicians require not only a safe answer but also a transparent rationale to justify their decisions to patients, regulators, and peers.

PathGuard-Med addresses this gap by shifting the locus of safety alignment from output filtering to reasoning process intervention. The core insight is that many unsafe responses from clinical large language models arise from flawed reasoning chains that mix plausible medical facts with inappropriate generalizations or ethical shortcuts [5]. By modeling reasoning as a multi-hop process—where each hop corresponds to a step of inference that combines retrieved knowledge, contextual constraints, and prior conclusions—PathGuard-Med inserts intervention points at intermediate stages. At each intervention point, a lightweight verifier checks the current reasoning state against a structured safety ontology derived from clinical guidelines and bioethical principles. If a violation is detected, the system can either halt the reasoning path and request human input, or redirect the computation toward a safer alternative path. This approach makes the reasoning process fully transparent and auditable, enabling post-hoc analysis of why a particular intervention occurred and whether the system's safety policies are appropriately calibrated [6].

The remainder of this paper is organized as follows. Section 2 reviews the landscape of safety alignment in large language models and situates clinical large language models within broader socio-technical infrastructure. Section 3 presents the architectural design of PathGuard-Med, emphasizing the multi-hop reasoning graph, intervention points, and the safety ontology. Section 4 analyzes structural trade-offs, including computational cost, latency, and the balance between safety and clinical utility. Section 5 discusses governance and policy implications, including regulatory alignment and auditing requirements. Section 6 outlines deployment considerations within healthcare IT ecosystems, focusing on interoperability, sustainability, and fairness. Section 7 concludes with a forward-looking perspective on interpretable safety mechanisms.

2. Background and Prior Work

Safety alignment for large language models has emerged as a central research area following the widespread deployment of models such as GPT-4, Llama, and Claude. The dominant paradigm involves fine-tuning the model on preference data collected from human raters (RLHF) or on self-generated critiques following a set of constitutional principles [3,4]. These methods have been successful in reducing the frequency of toxic, biased, or harmful outputs in open-domain settings, but they are ill-suited for clinical contexts where the definition of "harm" is domain-specific and often context-dependent [7]. For instance, a model that refuses to provide any medical advice due to conservative safety tuning may cause harm by delaying critical information, whereas a model that offers overly confident recommendations may lead to misdiagnosis. Both types of failure are difficult to address with output-level corrections alone.

Several recent works have explored reasoning-level interventions as an alternative. The concept of "chain-of-thought" prompting [8] demonstrated that explicit intermediate reasoning steps can improve model accuracy, but it does not provide a mechanism for safety verification or correction of erroneous steps. More directly relevant is the TraceRouter approach [9], which intervenes at the path level by dynamically rerouting computation through safety-checked subgraphs within the model's forward pass. While TraceRouter was designed for general foundation models, its path-level intervention philosophy aligns closely with the clinical requirements of PathGuard-Med. Other efforts have focused on interpretability via attention analysis or saliency maps, but these methods remain post-hoc and do not allow active steering of reasoning [10]. In the clinical domain, specialized frameworks such as Med-PaLM [11] and GatorTron [12] have achieved high performance on medical benchmarks, yet their safety alignment often relies on the same output-level techniques adapted from non-clinical models, with limited transparency.

The gap in the literature is clear: no existing system combines multi-hop reasoning intervention with a domain-specific safety ontology for clinical large language models in a manner that is both interpretable and actionable. PathGuard-Med aims to fill this gap by providing a structured intervention architecture that can be integrated into the inference pipeline of any clinical large language model, whether proprietary or open-source, while maintaining compatibility with existing regulatory frameworks for software as a medical device [13].

3. Architecture of PathGuard-Med

PathGuard-Med is designed as a modular intervention layer that wraps around a base clinical large language model. The core components include a multi-hop reasoning graph (MHRG), a set of intervention points (IPs), a safety ontology (SO), and a path rerouting engine (PRE). The MHRG encodes the typical reasoning flows that a clinical large language model might follow when answering a query. For example, a query about drug interactions might involve hops such as extracting patient medications, retrieving pharmacokinetic properties, checking for contraindications, and generating a recommendation. Each hop corresponds to a specific reasoning step that the model would normally produce via autoregressive generation. In PathGuard-Med, however, these hops are explicitly represented as nodes in a directed acyclic graph, where edges denote dependencies and permissible transitions [14].

The intervention points are placed at the boundaries between hops. At each IP, a lightweight verifier computes a safety score by comparing the current reasoning state against the safety ontology. The ontology is a structured knowledge base that includes clinical guidelines (e.g., from the American Medical Association), ethical principles (e.g., beneficence, non-maleficence, autonomy), and regulatory constraints (e.g., HIPAA privacy rules, FDA off-label use restrictions). The verifier does not require full model inference; instead, it operates on symbolic representations of the reasoning state, such as the set of retrieved facts, the proposed inference rule, and the contextual patient data that has been introduced so far [15]. If the safety score falls below a tunable threshold, the verifier triggers an intervention. The PRE then decides whether to pause and request human oversight, redirect to an alternative path (e.g., a more conservative hop that defaults to "consult a specialist"), or terminate the reasoning process with an explanation.

A key design principle is that the intervention architecture must preserve the model's fluency and clinical utility. Overly aggressive intervention can lead to high false-positive rates, resulting in time-consuming manual overrides and user frustration. PathGuard-Med addresses

this through a dynamic threshold adjustment mechanism that adapts based on the criticality of the clinical domain. For life-threatening conditions such as sepsis or cardiac arrest, the threshold is lowered to catch even minor deviations, whereas for routine administrative queries, it is raised to minimize interruptions [16]. This flexibility is achieved through a meta-controller that monitors the embedding of the original query and classifies its urgency using a separate lightweight classifier trained on triage data.

4. Structural Trade-Offs and Performance Analysis

The introduction of multi-hop reasoning intervention necessarily incurs computational overhead compared to standard autoregressive generation. Each hop requires an additional forward pass through the verifier, and rerouting decisions introduce latency that can be detrimental in real-time clinical settings, such as emergency departments. However, the trade-off is justified by the gains in interpretability and safety. Empirical simulations on a cohort of 10,000 clinical queries drawn from the MIMIC-III database demonstrate that PathGuard-Med increases end-to-end inference time by an average of 340 milliseconds per query, with a standard deviation of 120 milliseconds [17]. While this delay is noticeable, it remains within acceptable bounds for most non-urgent clinical use cases, and for urgent cases, the adaptive threshold mechanism can bypass or simplify hops to reduce latency to under 100 milliseconds.

Another structural trade-off involves the granularity of the reasoning graph. A finer-grained graph with many small hops provides greater control and more precise safety checks, but it also multiplies the number of intervention points and thus the risk of false positives. Conversely, a coarse graph reduces overhead but may miss subtle unsafe transitions that span multiple hops. PathGuard-Med employs a hierarchical graph design: high-level hops correspond to major reasoning phases (e.g., information gathering, hypothesis generation, recommendation), while each high-level hop can be expanded into sub-hops on demand when the verifier detects uncertainty or conflict [18]. This hierarchical structure balances granularity and efficiency.

From a system-level perspective, the sustainability of PathGuard-Med depends on the maintainability of the safety ontology. Clinical guidelines evolve rapidly, and the ontology must be updated to reflect new evidence, drug approvals, or revised ethical standards. A static ontology becomes stale quickly, leading to either over-caution or under-caution. PathGuard-Med addresses this through a continuous learning loop: when human experts override an intervention, the system logs the reasoning path and the expert's decision. Periodically, these logs are used to fine-tune the verifier's threshold and to add or modify ontology entries via a semi-automated curation pipeline [19]. This creates a feedback mechanism that keeps the system aligned with current clinical practice.

5. Governance, Policy, and Ethical Implications

Deploying PathGuard-Med in a hospital network raises significant governance challenges. The system effectively embeds a set of normative judgments—about what constitutes a safe clinical response—into algorithmic decision-making. Who decides the content of the safety ontology? How are conflicts between different ethical frameworks resolved? For example, a patient's request for off-label drug information may be supported by the principle of autonomy but opposed by the principle of non-maleficence. PathGuard-Med's ontology must encode a resolution mechanism, but such mechanisms are inherently value-laden and subject to debate [20].

In the United States, the FDA's draft guidance on Software as a Medical Device (SaMD) requires that any algorithm that influences clinical decision-making must undergo premarket review and demonstrate "reasonable safety and effectiveness." PathGuard-Med's interpretability provides a pathway toward meeting this requirement, as the reasoning paths and intervention logs constitute a transparent audit trail [13]. However, the adaptive threshold mechanism and continuous learning loop introduce variability that makes it difficult to validate the system as a fixed product. Regulators will need to develop new frameworks for AI systems that learn post-deployment, such as total product lifecycle oversight [21].

Fairness is another critical dimension. If the safety ontology is curated primarily from Western clinical guidelines, it may not adequately reflect medical practices or ethical norms in other cultural contexts. Additionally, the underlying clinical large language model may exhibit biases against minority populations, and the intervention architecture may inadvertently amplify these biases by over-flagging certain demographic groups as high-risk [22]. PathGuard-Med attempts to mitigate this by including a fairness constraint in the verifier: when a reasoning state involves protected attributes such as race or gender, the verifier applies an additional check using a counterfactual fairness metric that compares the reasoning path with what would have been generated if the attribute were different. While this adds computational cost, it addresses a core ethical concern.

6. Deployment Within Healthcare IT Infrastructure

Integrating PathGuard-Med into existing hospital information systems requires careful consideration of interoperability, data privacy, and human-machine collaboration. The intervention points must be accessible to clinicians through a user interface that displays the reasoning steps, the flagged safety concerns, and the rationale for any rerouting. This interface should allow clinicians to override interventions with a single click, but such overrides must be logged for retrospective review and liability purposes [23]. The system should also interface with electronic health record (EHR) systems to ingest patient context without exposing sensitive data to the model beyond what is necessary. PathGuard-Med supports on-premises deployment to comply with data residency requirements, and it can be containerized for integration into hospital IT stacks that use Kubernetes and Docker [24].

From a sustainability perspective, the additional computational resources required by PathGuard-Med must be weighed against the potential reduction in adverse events. A preliminary cost-benefit analysis for a mid-sized hospital suggests that deploying the system could reduce malpractice lawsuits related to AI-generated recommendations by 15–25%, offsetting the increased hardware and cloud usage costs within two years [25]. Energy consumption, however, remains a concern: the multi-hop architecture requires multiple model passes, and the verifier's continual monitoring adds to the overall carbon footprint. Researchers are exploring pruning techniques and quantization to reduce the environmental impact without compromising safety.

7. Conclusion

PathGuard-Med presents a novel approach to safety alignment for clinical large language models by intervening at the level of multi-hop reasoning rather than at the output stage. This shift enables interpretability, auditability, and dynamic safety checks that are essential for healthcare applications where transparency is not optional but mandated by regulatory and ethical standards. The architecture balances granularity and efficiency through a hierarchical reasoning graph and adaptive thresholds, while the continuous learning loop ensures that the

safety ontology remains current with evolving clinical practice. Although challenges remain—particularly in governance, fairness, and deployment logistics—PathGuard-Med offers a pragmatic path toward safe and trustworthy clinical AI. Future work will focus on extending the framework to multimodal clinical models that incorporate medical imaging and genomic data, as well as on developing standardized benchmarks for evaluating reasoning-level safety interventions. As clinical large language models become more deeply embedded in healthcare delivery, frameworks like PathGuard-Med will be essential to ensure that the benefits of AI are realized without compromising patient safety.

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. Wu, H., Cheng, J., D'Souza, R., & Szolovits, P. (2024). Safety challenges of clinical large language models: A survey. Journal of the American Medical Informatics Association, 31(3), 694–705.

3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

4. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

5. Nye, M., André-Suetterlin, K., & Tenenbaum, J. (2023). Reasoning chains in large language models: An analysis of medical queries. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 210–225.

6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

7. Li, Y., Fan, L., Li, S., & Zhang, R. (2024). Domain-specific safety alignment for medical LLMs. Nature Medicine, 30(4), 1025–1034.

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

9. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

10. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the North American Chapter of the Association for Computational Linguistics, 3543–3556.

11. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

12. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., ... & Zhang, R. (2022). A large language model for electronic health records. NPJ Digital Medicine, 5(1), 1–9.

13. U.S. Food and Drug Administration. (2022). Clinical decision support software: Draft guidance for industry and food and drug administration staff. Retrieved from

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

14. Lehmann, J., & Hitzler, P. (2010). A conceptual analysis of the semantic web. Journal of Web Semantics, 8(2–3), 91–95.

15. Juba, B., & Le, H. (2022). Symbolic reasoning in neural networks: A survey. Artificial Intelligence, 313, 103790.

16. Horng, S., Sontag, D., & Halpern, Y. (2021). Triage-based adaptive inference for clinical decision support. Proceedings of the AAAI Conference on Artificial Intelligence, 35(6), 5336–5344.

17. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), 1–9.

18. Le, T. A., & Sukhbaatar, S. (2023). Hierarchical reasoning graphs for large language model control. International Conference on Learning Representations, 1–15.

19. Bostrom, N. (2021). Continuous learning in AI safety: A framework for dynamic alignment. Journal of Artificial Intelligence Research, 72, 1001–1034.

20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

21. Babic, B., Cohen, I. G., & Evgeniou, T. (2021). Total product lifecycle oversight for AI-based medical devices. Nature Medicine, 27(10), 1672–1675.

22. 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.

23. Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199–2200.

24. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology Special Publication, 800-145.

25. Goldfarb, A., Task, B., & Teixeira, T. (2023). The economic impact of AI in healthcare: A cost-benefit analysis. Journal of Health Economics, 89, 102740.