1. GroundedMed-LLM: Enhancing Clinical LLM Safety via Knowledge Graph and Causal Reasoning Integration

May 10, 2025

Abstract

Large Language Models (LLMs) show promise in healthcare but suffer from "hallucinations"-generating plausible yet incorrect information-and a lack of grounding in established medical knowledge and causal reasoning, posing risks to patient safety. This research addresses this critical gap by proposing "GroundedMed-LLM," a novel framework designed to enhance LLM safety and reliability for clinical support. Our approach involves deeply integrating LLMs with curated medical Knowledge Graphs (KGs) and explicit causal reasoning modules. The methodology includes developing robust KG-to-LLM integration mechanisms, such as dynamic knowledge retrieval and an iterative refinement loop driven by KG validation, and incorporating causal models to ensure outputs are clinically coherent and adhere to established cause-and-effect relationships. GroundedMed-LLM will be prototyped and rigorously evaluated on clinical support tasks like differential diagnosis, using metrics for factual accuracy, clinical validity, KG grounding, causal consistency, and aiming to demonstrate a quantifiable reduction in harmful hallucinations. We expect this framework to significantly improve LLM trustworthiness, increase clinician adoption, support better clini∏cal decision-making, and contribute to the development of responsible AI in healthcare.

Contents

0.1 1.1. Abstract

Large Language Models (LLMs) show promise in healthcare but suffer from "hallucinations"—generating plausible yet incorrect information—and a lack of grounding in established medical knowledge and causal reasoning, posing risks to patient safety. This research addresses this critical gap by proposing "GroundedMed-LLM," a novel framework designed to enhance LLM safety and reliability for clinical support. Our approach involves deeply integrating LLMs with curated medical Knowledge Graphs (KGs) and explicit causal reasoning modules. The methodology includes developing robust KG-to-LLM integration mechanisms, such as dynamic knowledge retrieval and an iterative refinement loop driven by KG validation, and incorporating causal models to ensure outputs are clinically coherent and adhere to established cause-and-effect relationships. GroundedMed-LLM will be prototyped and rigorously evaluated on clinical support tasks like differential diagnosis, using metrics for factual accuracy, clinical validity, KG grounding, causal consistency, and aiming to demonstrate a quantifiable reduction in harmful hallucinations. We expect this framework to significantly improve LLM trustworthiness, increase clinician adoption, support better clini∏cal decision-making, and contribute to the development of responsible AI in healthcare.

---

1. Background & Literature Review

Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs) based on the Transformer architecture (Vaswani et al., 2017), offers significant potential to revolutionize healthcare. LLMs like GPT variants (Brown et al., 2020) are being explored for diverse clinical applications, including automating clinical note generation, summarizing patient records, and developing medical chatbots (Liang et al., 2022; Lee et al., 2023). These models can process and generate human-like text, promising to alleviate administrative burdens and improve information access.

However, a critical barrier to the safe and effective deployment of LLMs in clinical settings is their propensity for "hallucinations"—generating information that is plausible-sounding but factually incorrect, clinically invalid, or not grounded in evidence (Singhal et al., 2023). This unreliability often stems from LLMs being trained on vast, general text corpora, lacking inherent mechanisms to prioritize or verify against established medical knowledge. Furthermore, standard LLMs often struggle with understanding complex medical causality, which is fundamental to clinical reasoning. The "black box" nature of many LLMs also hinders clinician trust and the ability to verify information, posing substantial risks if used in clinical decision support without robust safeguards.

Current strategies to enhance LLM reliability, such as fine-tuning on medical corpora or standard Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), offer partial improvements. RAG allows LLMs to retrieve information from external knowledge bases, but often falls short in ensuring deep semantic understanding, complex reasoning, or adherence to intricate clinical logic and causal constraints. Medical Knowledge Graphs (KGs) (e.g., UMLS, SNOMED CT) provide structured, verified medical knowledge, while causal reasoning principles are essential for understanding disease progression and treatment effects. There is a pressing need for frameworks that deeply integrate these elements to ensure LLM outputs are not only factually accurate but also clinically coherent and safe.

2. Problem Statement & Research Gap

3. 2.1. Problem Statement

Current Large Language Models (LLMs), despite their advanced natural language capabilities, are insufficiently reliable for many critical clinical support tasks. Their tendency to "hallucinate"—generating information that is factually incorrect, not grounded in evidence, or clinically unsafe—poses significant risks to patient safety and undermines clinician trust. This unreliability arises because LLMs often lack robust mechanisms to access, verify, and reason with established medical knowledge or understand causal relationships crucial to clinical contexts.

4. 2.2. Research Gap

A critical research gap exists in developing methodologies that enable LLMs not just to retrieve relevant facts from medical knowledge sources, but to actively reason with them and adhere to causal principles. While methods like basic RAG exist, they often do not ensure deep semantic grounding or the application of complex clinical logic. This research aims to bridge this gap by developing and evaluating "GroundedMed-LLM," an integrated framework that deeply synergizes medical KGs and explicit causal reasoning modules with LLMs to produce outputs that are demonstrably safer, more accurate, and clinically valid for healthcare applications.

---

5. Proposed Gen AI Approach (Methodology)

6. 3.1. Overall Research Aim

To develop and evaluate "GroundedMed-LLM," a novel framework that integrates medical Knowledge Graphs (KGs) and causal reasoning mechanisms with Large Language Models to significantly enhance their factual accuracy, clinical safety, and interpretability for clinical support tasks.

7. 3.2. Objective 1: Development of a KG-Integrated LLM Architecture

- **3.2.1. Medical Knowledge Graph Curation/Selection:** We will primarily adapt and extend existing comprehensive medical KGs like UMLS and DrugBank, focusing on areas such as drug interactions and common disease pathways relevant to the evaluation use cases. A highly focused, purpose-built KG will be a secondary consideration if existing resources prove inadequate. An efficient API for KG querying will be utilized.
- **3.2.2. KG-to-LLM Integration Mechanisms:**
- **Dynamic Retrieval Module:** This module will identify and retrieve relevant entities and relations from the KG based on the input query and LLM generation context, using techniques like entity linking and graph traversal.
- **Knowledge Injection Techniques:** Our primary approach for knowledge injection will be augmenting input prompts with structured KG facts (e.g., linearized triples). Should access to open-source models (e.g., Llama variants) permit, we will explore fine-tuning with adapter layers; otherwise, prompt engineering will be refined for API-based models (e.g., GPT-class).
- **Iterative Refinement Loop:** An LLM-generated initial response will be cross-referenced against the KG. Discrepancies will trigger a refinement step, prompting the LLM to revise its output for improved accuracy based on KG evidence.

8. 3.3. Objective 2: Incorporation of Causal Reasoning Modules

- **3.3.1. Causal Model Representation:** Initially, we will focus on a limited set of high-confidence causal links pertinent to a chosen use case (e.g., common drug interactions or symptom-disease relationships in internal medicine), extracted from established clinical guidelines and validated by expert review. This acknowledges the significant undertaking of comprehensive causal model creation. Representations may include causal rules or simplified causal graphs.
- **3.3.2. Causal Constraint and Inference Mechanisms:**
- **Causal Consistency Checker:** This module will verify LLM-generated statements against the defined causal model, flagging inconsistencies (e.g., suggesting a contraindicated drug).
- **Simple Causal Inference Engine:** Enables basic causal inferences based on input, KG information, and the causal model (e.g., prioritizing a disease in differential diagnosis if it commonly causes presented symptoms).

- **Causally-Guided Generation:** We will explore methods to use causal graph structures to guide the LLM's generation, for example, by incorporating causal constraints into the LLM's decoding process or by using the causal model to rank or re-score LLM-generated hypotheses, promoting clinically plausible reasoning.

9. 3.4. Objective 3: Prototyping and Evaluation of GroundedMed-LLM

- **3.4.1. Prototype Development:** A prototype of GroundedMed-LLM will be implemented focusing on 1-2 clinical support use cases (e.g., preliminary differential diagnosis, clinician query answering). An existing pre-trained LLM (e.g., an open-source Llama variant or a GPT-class model accessible via API) will serve as the base model. The choice will influence the feasibility of certain integration techniques.
- **3.4.2. Dataset Preparation:** We will utilize existing de-identified clinical datasets (e.g., MIMIC-IV (Johnson et al., 2023)) where possible. For tasks like differential diagnosis, synthetic or semi-synthetic datasets (e.g., case vignettes authored by medical experts or adapted from educational materials to ensure clinical plausibility and ground truth accuracy) will be developed with gold-standard answers for evaluation.
- **3.4.3. Evaluation Metrics:**
- **Factual Accuracy:** Percentage of correct statements; entity-level precision/recall against KG.
- **Clinical Validity/Safety:** Expert clinician review (Likert scales) for safety, appropriateness, utility; quantification of harmful hallucinations.
- **KG Grounding Score:** Proportion of factual claims traceable to KG evidence.
- **Causal Consistency Score:** Adherence to defined causal models (expert/automated checks).
- **Reduction in Hallucinations:** Comparative analysis against baselines.
- **Interpretability:** Qualitative assessment of system's ability to provide justifications.
- **3.4.4. Experimental Setup:** GroundedMed-LLM will be compared against baselines: (a) standard pre-trained LLM, (b) LLM fine-tuned on medical data, (c) LLM with standard RAG. Ablation studies will assess contributions of KG integration and causal reasoning modules.

---

10. Expected Impact in Healthcare

The GroundedMed-LLM framework is anticipated to yield significant positive impacts:

- **Enhanced LLM Safety, Reliability, and Trust:** By grounding outputs in verified medical KGs and ensuring causal consistency, the framework will reduce incorrect or harmful information, making LLMs safer and fostering greater clinician trust and adoption. This directly addresses the critical need for reliable AI in clinical support.
- **Improved Clinical Decision-Making Support:** A more dependable AI assistant for tasks like information synthesis, differential diagnosis exploration, and guideline adherence checks can lead to more informed clinical decisions and potentially better patient outcomes.
- **Foundation for Advanced and Responsible AI:** This research will provide a methodology and architectural blueprint for developing more trustworthy generative AI systems in healthcare. It contributes to responsible AI innovation by prioritizing safety and alignment with clinical realities.

---

11. Limitations or Ethical Considerations

12. 5.1. Limitations of the Proposed Research

- **KG/Causal Model Scope:** Performance depends on the comprehensiveness, accuracy, and currency of the integrated KG and causal models. Initial models will necessarily be limited in scope.
- **Complexity of Medical Causality:** Capturing the full nuance of medical causality is a long-term AI challenge; this work will use simplified models.
- **Scalability & Generalizability:** Computational intensity and generalizability to diverse clinical areas or LLMs will require further investigation beyond initial use cases.
- **Bias in Knowledge Sources:** KGs and causal models may inherit biases from their source data; this is an ongoing challenge.

13. 5.2. Ethical Considerations

This research will proactively address critical ethical issues:

- **Bias and Fairness:**
- *Consideration:* KGs and causal rules may reflect existing biases.

- *Mitigation:* Audit knowledge sources for potential biases. Explore mitigation or flagging strategies. Evaluate fairness across subgroups if data permits.

- **Accountability and Responsibility:**

- *Consideration:* AI's role in adverse outcomes.

- *Mitigation:* Position GroundedMed-LLM as a clinical *support* tool, with clinicians retaining ultimate decision-making responsibility. Document system limitations and the necessity of human oversight.

- **Explainability and Interpretability:**

- *Consideration:* "Black box" nature of LLMs.

- *Mitigation:* Enhance interpretability by linking outputs to KG facts and causal rules. Achievable explainability will be an evaluation metric.

- **Data Privacy and Security:**

- *Consideration:* Use of sensitive patient data.

- *Mitigation:* If patient data is used (e.g., MIMIC-IV), it will be strictly de-identified and handled per regulations (e.g., HIPAA) and IRB approvals. Explore synthetic data to minimize real data use.

- **Potential for Misuse & Over-Reliance:**

- *Consideration:* Misinterpretation or deskilling.

- *Mitigation:* Design as an assistive tool augmenting clinical judgment. Clearly communicate intended use, capabilities, and limitations to encourage critical assessment.

- **Informed Consent:**

- *Consideration:* For data use or clinician participation.

- *Mitigation:* Secure IRB approval and appropriate informed consent for any research stages involving human participants or non-public patient data.

---

14. References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

- Johnson, A. E. W., Bulgarelli, L., Pollard, T. J., Horng, S., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV (version 2.2). *PhysioNet*. https://doi.org/10.13026/6mm1-ek60.

- Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233-1239.
- 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.
- Liang, H., Tsou, B. K., & Wang, W. (2022). A survey on clinical natural language processing. *Journal of Biomedical Informatics*, 135, 104223.
- 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.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Zakka, K., Ghantasala, K., Rawat, A.S., & Zaheer, M. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*.

Multi-Agent Reflection

> *The development of this research proposal benefited significantly from a simulated multi-agent collaborative approach, where different roles contributed to shaping the final document: 1.* ***Researcher Agent:*** *This persona initiated the core idea—addressing LLM unreliability in healthcare—drawing upon an initial literature review. It identified the potential of integrating Knowledge Graphs (KGs) and causal reasoning as a novel solution. This agent laid the groundwork by defining the problem and outlining a preliminary set of objectives and methods. 2.* ***Domain Expert Agent (Healthcare):*** *This perspective, informed by the "Domain Validation" report, emphasized the clinical relevance and practical implications of the proposed research directions. It highlighted the importance of multimodal data, proactive health, personalized digital twins, and critically, the foundational need for standardized evaluation metrics and rigorous clinical validation. This input helped prioritize aspects like clinical validity in the evaluation metrics and ensured the "Expected Impact" section resonated with real-world healthcare needs. It also reinforced the ethical considerations concerning patient safety and data privacy. 3.* ***Critic Agent:*** *This agent provided a crucial review of the initial proposal draft. Its feedback was instrumental in identifying key weaknesses, such as excessive length, lack of methodological focus in certain areas, and the need for more decisive language. The critique pushed for sharpening the novelty, clarifying feasibility (e.g., KG sourcing, LLM access implications), and ensuring a pragmatic scope for complex tasks like causal model creation. The call for*

substantial condensation was a primary driver for the refinement process. 4. ***Writer Agent (Scientific Proposal Writer):*** *This persona focused on structuring the proposal according to academic standards, ensuring clarity, conciseness, and logical flow. It translated the critiques and expert insights into polished text, adhering to formatting requirements (like Markdown and section structure) and word count guidelines (e.g., for the abstract). This agent was responsible for the iterative process of drafting, receiving feedback, and refining the language to be persuasive and academically sound, ultimately aiming for a submission-ready document. The interplay between these agents was vital. The Researcher initiated, the Domain Expert grounded the work in clinical reality, the Critic identified flaws and areas for improvement, and the Writer synthesized these inputs into a coherent and compelling narrative. This iterative, multi-perspective approach led to a more robust, focused, and well-argued proposal than any single agent could have produced in isolation. For instance, the Critic's demand for conciseness forced the Writer and Researcher to prioritize information ruthlessly, while the Domain Expert's validation ensured that clinically relevant aspects were not lost in this process. This collaborative refinement significantly enhanced the overall quality and potential impact of the research proposal.*