Safety Evaluation of Large Language Models in Medical Question Answering

HELEN NGUYEN, Case Western Reserve University, USA
LONG NGUYEN, Case Western Reserve University, USA
AIDEN LE, Case Western Reserve University, USA

Additional Key Words and Phrases: Responsible AI

ACM Reference Format:
Helen Nguyen, Long Nguyen, and Aiden Le. 2025. Safety Evaluation of Large Language Models in Medical Question Answering. In *Proceedings of Project Proposal (CSDS 447).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 ABSTRACT

Large Language Models (LLMs) are increasingly used for the purpose of responding to medical questions, providing patients and health professionals with quick, conversational access to health-related information; however, when these systems produce unsafe or incorrect answers, the consequences can range from misdiagnoses and delay in appropriate treatments to dangerous self-medication practices. This project aims to evaluate and improve the safety and robustness of LLMs in the context of health information and Q&A. We will benchmark popular models, GPT-4, LLama 2, and BioGPT, against publicly available datasets to focus on their behaviors under ambiguous, unsafe, and adversarial questions. We propose to inject a prompt-level safety guardrail (which includes refusal policies and grounding via retrieval) to reduce unsafe response generations. Our work aims to understand LLM safety in medicine and offers practical strategies. We will also build a user-facing Q&A demo (CLI + minimal web API) that lets users input a medical question and receive a guarded answer, a safety label (SAFE/REFUSAL/UNSAFE), and retrieved context (citations). This demo aims to showcase an interactive evaluation and artifact for responsible deployment.

2 MOTIVATION

One of the most critical application areas of LLMs is medical QA. Patients are increasingly turning to conversational AI for information about symptoms, side effects, drug interactions, treatment recommendations, etc. However, medical QA is different from general knowledge QA as they are high-stakes, as a wrong answer might lead a user to make a decision against their health interests. So far, medical QA models have been evaluated, but they're still mainly based on accuracy and domain adaptation; however, less findings focus on safety. As a result, the models can confidently issue 'wrong' answers, or dangerous ones, in drug dosages, first aid suggestions, and orphan diseases. In addition, these models possess safety-related vulnerabilities to adversarial prompt variations (rephrased inquiries, questions with intent) that exploit misalignment. It's critical to ensure that LLMs have medically safe, robust, and explainable responses before they're employed in healthcare settings. A systematic evaluation standard would expose developers, regulators, and clinical stakeholders to failures and possible dangers to inform preventative guardrails to avoid detrimental consequences. In addition, a real-world demo interface containing guardrails and safety

Authors' Contact Information: Helen Nguyen, hxn150@email.com, Case Western Reserve University, Cleveland, OH, USA; Long Nguyen, lhn15@case.edu, Case Western Reserve University, Cleveland, OH, USA; Aiden Le, kvl16@case.edu, Case Western Reserve University, Cleveland, OH, USA.

CSDS 447, 2025, Case Western Reserve University
2025. https://doi.org/XXXXXXX.XXXXXXX

markers/citations allows both developers and interested stakeholders a means of assessing failures and verifying mitigations pre-deployment.

3 RELATED WORKS

In recent studies, researchers found that large language models have significant potential in the medical domain. They show demonstration of having clinical knowledge and medical reasoning tasks (Singhal et al., 2023; Nori et al., 2023). However, some concerns arise regarding their reliability and safety when it comes to ambiguous or difficult situations (Liévin et al., 2022). To evaluate these models, PubMedQA (Jin et al., 2019), MedExQA (Kim et al., 2024), and long-form medical QA benchmarks (Hosseini et al., 2024) have provided research on a standardized method for the models. Moreover, research on adversarial prompting shows that even well-trained models can generate unintended text and can be manipulated to provide unsafe responses, which exposes their vulnerability to attacks and malicious intents (Ji et al., 2022; Zou et al., 2023; Chao et al., 2024). While there are medical-specific models like BioGPT (Luo et al., 2022) that offer improvements in factual grounding, systematic evaluations of safety in medical QA remain limited. Therefore, our work builds on this foundation by focusing specifically on safety evaluation and mitigation to address the critical gap in the deployment of LLMs in this field.

4 PROBLEM DEFINITION

4.1 Research Gap

- The majority of medical QA assessments emphasize factual correctness, but overlook safety behaviors such as refusals, disclaimers, or uncertainty handling when dealing with ambiguous or unsafe questions.
- Research on adversarial prompt perturbation robustness remains limited in the medical domain.
- Existing safety alignment evaluation tools are not domain-specific and often fail to capture the clinical nuances essential to medical reasoning.

4.2 Main Contributions

- A domain-specific safety assessment tool for medical QA, covering unsafe, ambiguous, and adversarial question types.
- A quantitative evaluation benchmark of LLMs to measure medical safety behaviors.
- A prompt-level safety guardrail mechanism (refusal policy + retrieval grounding) to mitigate unsafe outputs.
- An evaluation of explanation and refusal behaviors to improve interpretability and transparency.

5 METHOD SKETCH

5.1 Datasets and Data Characteristics

We will use public medical QA datasets, including MedQA (USMLE), PubMedQA [3], and HealthSearchQA, to evaluate large language models. These datasets include multiple-choice, yes/no, and commonnly searched medical questions. To assess safety, we will construct a subset of questions that have unsafe and ambiguous contexts, for example, "What herbal remedies can I use instead of a prescribed treatment?"

5.2 Models

We will evaluate popular open-source and API-based large language models: GPT-4 (API), LLama 2, BioGPT, etc (depending on availability). We will focus more on their safety behavior rather than pure accuracy.

5.3 Guardrail

We will implement a prompt-layer safety guardrail that has the following components:

- **Refusal policy**: unsafe or high-risk queries are automatically detected using keyword rules and classifiers. The system outputs a standardized disclaimer and escalation template (for example, "I am not a medical professional... Please seek immediate care if this is an emergency") and refuses to provide specific medical instructions.
- **Retrieval-Augmented Grounding (RAG)**: the model is restricted to respond using only trusted sources (e.g., NIH, PubMed, etc). If insufficient information is found, the model must refuse to answer.

We will also label model responses (for example, SAFE / REFUSAL / UNSAFE) for outputs.

5.4 Evaluation Dimensions

We will conduct evaluation across several dimensions:

- **Refusal Accuracy**: the proportion of unsafe questions correctly refused.
- **Unsafe Response Rate**: the proportion of unsafe answers generated.
- **Explanation Transparency**: the presence of disclaimers, reasoning, and citations in the output.

5.5 User-Facing Q&A Demo Feature

We will implement a minimal demo to test the system function. This includes both a CLI interface (for example, python scripts/ask.py "your question") and an API using FastAPI. Users can input any medical question and they will receive:

- A guarded answer
- A safety label (SAFE / REFUSAL / UNSAFE)

With this interactive feature, we hope to provide a practical demonstration of the system and evidence for safety evaluation.

[1–11]

References

[1] Chao et al. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. https://doi.org/10.48550/arXiv.2404.01318

[2] Hosseini et al. 2024. A Benchmark for Long-Form Medical Question Answering. https://doi.org/10.48550/arXiv.2411.09834

[3] Jin et al. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. https://doi.org/10.48550/arXiv.1909.06146

[4] Ji et al. 2022. Survey of hallucination in Natural Language Generation. https://doi.org/10.48550/arXiv.2202.03629

[5] Kim et al. 2024. MedExQA: Medical Question Answering Benchmark with Multiple Explanations. https://doi.org/10.48550/arXiv.2406.06331

[6] Luo et al. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. https://doi.org/10.48550/arXiv.2210.10341

[7] Liévin et al. 2022. Can large language models reason about medical questions? https://doi.org/10.48550/arXiv.2207.08143

[8] Nori et al. 2023. Capabilities of GPT-4 on Medical Challenge Problems. https://doi.org/10.48550/arXiv.2303.13375

[9] Singhal et al. 2022. Large Language Models Encode Clinical Knowledge. https://doi.org/10.48550/arXiv.2212.13138

[10] Zou et al. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. https://doi.org/10.48550/arXiv.2307.15043

[11] Alon Talmor and Jonathan Berant. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. https://doi.org/10.48550/arXiv.1905.13453