Research proposal: Output correctness guarantees with LLM guardrails

Yasin Gül

Arjan Broer
Open University of the Netherlands

Abstract

Large Language Models (LLMs) demonstrate impressive performance in a wide range of tasks, yet their outputs are not always correct or reliable. This introduces challenges in applications where factual or logical correctness is critical. This research investigates whether guardrails—mechanisms designed to constrain or verify model output—can provide guarantees for output correctness. Using a design research approach, we will iteratively develop and evaluate such guardrails, focusing on how correctness can be defined, implemented, and assessed. The study aims to identify what types of guarantees are feasible and what trade-offs are involved in enforcing them, with a particular focus on medical queries about medications.

Background

Large Language Models (LLMs) are capable of generating coherent and contextually appropriate responses. However, they occasionally produce hallucinations, factual inaccuracies, or unsafe advice. This risk becomes particularly critical in medical contexts, where incorrect information can have serious consequences.

Examples include suggesting dangerous dosages, recommending inappropriate medications, or failing to warn against contraindications. For instance, providing a baby with an adult dose of ibuprofen, prescribing antibiotics for viral infections, or advising unsafe use of controlled substances are all examples where hallucinated outputs could cause harm.

Guardrails have emerged as a strategy to mitigate these risks by constraining or verifying model outputs. Techniques such as retrieval-augmented generation (RAG) and output moderation models have been proposed to systematically improve output reliability (**?**). Retrieval-augmented generation grounds answers in external, verified sources, reducing hallucinations, while moderator models can filter unsafe or incorrect content after generation (**?**).

In the medical domain, ensuring correct advice is even more crucial, particularly regarding medications. Recent work shows that LLMs can contribute to reducing medication instruction errors when proper safeguards are in place (**?**).

This proposal explores whether guardrails can be systematically designed and combined to guarantee correctness in LLM-generated answers for medication-related medical prompts. We particularly investigate the effectiveness of retrieval-augmented generation and moderator models, individually and in combination, to prevent incorrect or unsafe medical advice.

Research questions

This research explores the design and evaluation of guardrails that aim to ensure the correctness of outputs produced by LLMs for medication-related medical queries. The central questions guiding this work are:

RQ1. Can a guardrail mechanism combining RAG and moderator models ensure 100% correctness in LLM responses to medical prompts?

RQ2. How can correctness be operationalized for medication-related queries (e.g., dosage, indications, contraindications)?

RQ3. What are the strengths and limitations of using RAG and moderator models in ensuring factual correctness in the medical domain?

Research methods

This project will follow a design research methodology, consisting of iterative cycles of building, testing, and refining artifacts. The artifact in this case is a guardrail mechanism for LLMs that constrains or validates generated output.

We will test and compare four experimental conditions to evaluate the effectiveness of different guardrail approaches for ensuring correctness in medical prompts:

- **Baseline:** No guardrails applied.
- **RAG-only:** Responses are grounded in external verified medical sources using retrieval-augmented generation (**?**).
- **Moderator-only:** A second AI model evaluates the generated output and flags or blocks incorrect or unsafe content (**?**).
- **Combined:** Both RAG and moderator-AI are applied sequentially—first grounding the answer, then checking its correctness.

The evaluation focuses on whether the guardrail meets defined correctness criteria in realistic scenarios and what design choices contribute to its effectiveness.

Apparatus

The exact technical setup for this project has not yet been determined. An initial idea is to use existing large language model agents, such as ChatGPT with function-calling capabilities, to simulate retrieval-augmented generation and moderation behaviors without requiring full custom implementation in Python.

If this approach proves insufficient for the intended experiments, alternative methods will be explored. Possible options include low-code or API-based solutions, or minimal programming to construct basic prototypes.

The final choice of tools and models will depend on feasibility, ease of integration, and their suitability for testing correctness guardrails in medical prompts.

Stimuli

To evaluate the guardrails, we will use a fixed set of prompts that represent high-risk medical questions, particularly about medications (e.g., dosage, indications, and contraindications). These tasks represent real-world LLM use cases with identifiable correctness constraints (**?**).

Design

Each condition will be tested across the same prompt set. Performance will be evaluated in terms of correctness, safety, and response quality. The design is iterative, allowing refinements to each method across evaluation cycles.

Procedure

1. Identify output correctness criteria for selected medical prompts.
2. Implement guardrail mechanisms: RAG, moderator model, and their combination.
3. For each prompt, generate model outputs in four conditions: Baseline, RAG-only, Moderator-only, and Combined.
4. Compare outputs against correctness criteria using expert annotation or rule-based validation.
5. Analyze failures, compare effectiveness, and refine guardrail designs.

Data analysis

Each guardrail condition (Baseline, RAG-only, Moderator-only, Combined) will be evaluated using the following metrics:

- **Error rate:** Percentage of incorrect or unsafe answers.
- **Precision/Recall:** For detecting unsafe or medically incorrect outputs.
- **False positives / negatives:** Analysis of over- or under-blocking.
- **Qualitative analysis:** Categorization of failure types per condition.
- **Reflective analysis:** Documentation of design choices and their impact.

This comparison allows us to determine whether a specific approach—or their combination—can achieve 100% correctness or significantly reduce critical errors.

Time schedule

The research will be conducted as part of a course. The proposal for the research will be submitted on 30 April 2025. A peer review of the proposal will be conducted by a fellow student. Feedback on the proposal will be provided shortly after 15 May 2025. The final report will be submitted on 15 July 2025.