AI TRUSTWORTHINESS USING RETRIEVAL AUGMENTED GENERATION FOR POLITE PATIENT PRESCRIPTION EXPLANATORY TEXTS

Robert Devenyi
Carnegie Mellon University
Pittsburgh, PA 15213
rdevenyi@andrew.cmu.edu

Haochen Yang
Carnegie Mellon University
Pittsburgh, PA 15213
hy3@andrew.cmu.edu

John Turner-Smith
Carnegie Mellon University
Pittsburgh, PA 15213
jturners@andrew.cmu.edu

Hayden Stec*
Carnegie Mellon University
Pittsburgh, PA 15213
hstec@andrew.cmu.edu

September 17, 2023

ABSTRACT

The potential risks and rewards of introducing black box foundation models into medical interventions are great. Care must be taken in considering which interventions are appropriate for these models and how to best ensure patient safety and trust in the system. We utilize retrieval augmented generation to prompt large language models with factually correct information supported by an external database. Our application in generating polite, patient-forward text demonstrates a concrete and reliable step in trustworthy, transparent application of large language models for explaining medications to patients.

Keywords LLMs · GPT-4 · Foundation Models · AI Trust · Patient Safety · Retrieval Augmented Generation · Prescription Medication

1 Introduction

Since the COVID-19 pandemic, trust in medical doctors, scientists, and the healthcare industry has declines significantly [Kennedy et al., 2022]. Trust in medicine is difficult to rebuild, however small steps can be made using safe, trustworthy technology that can enhance patients' healthcare experience. Often in online healthcare management systems, multiple touch-points exist to connect patients with healthcare professionals, but difficulties in communication, assumptions regarding prior knowledge, and more can negatively impact patients' experience, and in the worst case lead to malpractice [Humphrey et al., 2022]. Foundation models such as GPT-4 have already proven the ability to perform on standardized medical examinations and show broad medical domain knowledge OpenAI, 2023. In this paper, we introduce DRAI, an implementation of retrieval augmented generation (RAG) to improve upon GPT-4's capabilities in generating patient-froward explanatory text about new diagnoses, prescriptions, and other additions to a patient's medical record. We also examine DRAI in a comparison against a non-RAG prompting of GPT-4 and evaluate the two for a number of metrics.

2 Background

AI Safety and Trust. LLMs and other foundation models, despite their power and broad applicability, are prone to *hallucinations*, or generation of false 'facts' and use of faulty reasoning [OpenAI, 2023, Li et al., 2023a, Chen

*Report produced with the support of Auton Lab for their 30th Anniversary "hackAuton" hackathon

et al., 2023]. Hallucinations are likely to degrade trust in AI systems, but RAG has had promising results in improving accuracy and transparency of texts [Li et al., 2023a, Lewis et al., 2021, Chen et al., 2023]. These recent developments provide the opportunity to expand usage of foundation models into the medical field in a manner conscious of the need for maintenance of patient trust in healthcare. In the paragraph below, we describe how these developments might impact medicine.

Patient Safety. Trust is a foundational component of the patient-doctor/healthcare relationship, and its decrease results in decreased health outcomes [Birkhäuer et al., 2017, Pearson and Raeke, 2000]. With advances in RAG, and overall improvements in foundation models, these tools might now be able to be implemented into medical apparatuses while still maintaining or perhaps improving general trust in medicine. Prior work using LLMs have found that LLMs without modifications such as fine tuning under-perform in specialized domains like medicine [Gutierrez et al., 2022]. Trust is also maintained by interpersonal politeness and clarity, which doctors often struggle to find a balance between [Aronsson and Sätterlund-Larsson, 1987]. However, recent work in the domain of autonomous vehicles finds that these interpersonal metrics can also apply to AI agents [gil Lee and Lee, 2022]. With this in mind, our work hopes to extend this analysis by examining whether an LLM given prompts with factual data confirmed by a pre-existing database can outperform the standard model without retrieval augmentation in accuracy, as well as politeness and clarity.

3 Methods

Our work leverages Open AI's GPT-4 foundation model via the Open AI API. As a proof of concept, our work relied on data from the openFDA database, a freely available API for products approved by the FDA.[1] First, we used openFDA to randomly generate a list of prescription medications for testing. Then, we generated two prompts for each prescription medication: one our benchmark condition and the other our experimental condition. Our benchmark consisted of prompting GPT-4 with the name of the prescription medication and that it should respond with an explanation of the medication relevant for a hypothetical patient. The experimental condition consisted of a similar prompt, but with the inclusion of relevant data-points from the openFDA entry for that prescription medication. We then prompted GPT-4 to include these facts. We then iterated across these data in a blind comparison on correctness, politeness, and clarity.

First, we randomly selected a diverse set of prescription medications using the openFDA database, aiming to cover a variety of medical conditions and treatment types.

Prompt Generation: For each selected medication, we constructed two distinct prompts for interaction with the GPT-4 model, representing benchmark and experimental conditions.

- Benchmark Condition Prompt: This prompt gave only the name of the prescription medication and instructed GPT-4 to generate an explanation relevant to patients.
- Experimental Condition Prompt: This prompt included not only the medication name but also integrated pertinent data points sourced from the openFDA entry for the corresponding prescription medication. GPT-4 was directed to incorporate these factual data into its response. This integration of openFDA data is aimed at retrieval generated augmentation of the prompt itself, ensuring verified and factual information about the prescription medication, its usual dosage and warnings, and more.

After presenting the prompts, we collected the generated text and organized it according to which medication it was in response. A blind comparative analysis was conducted to assess the correctness, politeness, and clarity of the responses generated by GPT-4 under the benchmark and experimental conditions.

4 Results

Preliminary results indicate a general increase in accuracy. DRAI implementation of RAG on GPT-4 outperforms the benchmark GPT-4, but with increased accuracy there also resulted a lack of clarity for the patient. Data was presented to the user that was not necessarily relevant, and the openFDA data it was drawing upon appears to contain acronyms and other technical terms that DRAI would use without explaining. As a result, although the benchmark GPT-4 was less accurate, and even prone to hallucinations, it performed better on clarity. Neither DRAI nor GPT-4 performed significantly higher than the other on politeness.

[1] `https://open.fda.gov/`

5 Discussion

DRAI ultimately performed well according to the primary goal of achieving accuracy and minimizing hallucinations.

With this initial work complete, our team hopes to proceed to build upon this system. Future work might integrate a patient's medical history as a consideration for the generation of explanations, drawing connections between pre-existing conditions and listed side effects and warnings. Additionally, prior work regarding incorporation of research papers through directional stimulus prompting would allow for explanatory texts to continually integrate new information, and perhaps even provide citations [Li et al., 2023b]. As DRAI becomes more robust, we hope to integrate this or similar systems into patient-facing portions of electronic healthcare management systems, allowing patients to easily access trustworthy, patient, and clear explanations of new prescriptions and diagnoses.

6 Acknowledgements

This work was supported by Auton Lab during hackAuton 2023, generously supported by Andrew Moore and Mary Soon Lee, Carnegie Mellon University, Pittsburgh Regional Health Initiative, Globodon, Edwards, Auton Systems, Kan Deng, Marinus Analytics, Madalina Fiterau, DE Shaw and Co., Devon and David Pablo Cohn, and the Patient Safety Technology Challenge. Thanks to Auton Lab and all the hackathon sponsors for providing the space, mentorship, and resources needed for this project.

References

Brian Kennedy, Alec Tyson, and Cary Funk. Americans' trust in scientists, other groups declines. Technical report, Pew Research Center, 2022.

Kate Humphrey, Melissa Sundberg, Carly E. Milliren, Dionne A. Graham, and Christopher P. Landrigan. Frequency and nature of communication and handoff failures in medical malpractice claims. *Journal of Patient Safety*, 2022.

OpenAI. Gpt-4 technical report, 2023.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. Trac: Trustworthy retrieval augmented chatbot, 2023a.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Johanna Birkhäuer, Jens Gaab, Joe Kossowsky, Sebastian Hasler, Peter Krummenacher, Christoph Werner, and Heike Gerger. Trust in the health care professional and health outcome: A meta-analysis. *PloS one*, 12(2):e0170988, 2017.

Steven D Pearson and Lisa H Raeke. Patients' trust in physicians: many theories, few measures, and little data. *Journal of general internal medicine*, 15:509–513, 2000.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL `https://api.semanticscholar.org/CorpusID:247475981`.

Karin Aronsson and Ullabeth Sätterlund-Larsson. Politeness strategies and doctor-patient communication. on the social choreography of collaborative thinking. *Journal of Language and Social Psychology*, 6(1):1–27, 1987.

Jae gil Lee and Kwan Min Lee. Polite speech strategies and their impact on drivers' trust in autonomous vehicles. *Computers in Human Behavior*, 127:107015, 2022. ISSN 0747-5632. doi:https://doi.org/10.1016/j.chb.2021.107015. URL `https://www.sciencedirect.com/science/article/pii/S0747563221003381`.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. Guiding large language models via directional stimulus prompting, 2023b.