RTHH: Robby the Hallucination Hunter - Evaluating and Reducing Medical Hallucinations in Large Language Models Using Structured Prompting

Arturo G. Valle

California Baptist University

arturog.valle@calbaptist.edu

Abstract

Large Language Models (LLMs) are now commonly used to answer health related questions, but they often produce hallucination information that sounds confident but is incorrect. This is a safety risk in medical situations. This project presents RTHH (Robby the Hallucination Hunter), a two-phase evaluation framework designed to measure and reduce hallucinations in medical question answering tasks.

In Phase 1, Gemini and ChatGPT were tested with a simple prompt using the MedQuad dataset. In Phase 2, the same questions were asked again using structured prompts that encouraged factual answers and recognized ambiguity. A transformer based embedding model was used to compare the model answers to the ground truth answers. The results show that structured prompting reduces hallucinations and increases accuracy for both models. This suggests that prompt engineering is a practical way to improve reliability without retraining the models.

Research Question: How do Gemini and ChatGPT respond to medical questions when using a simple baseline prompt, and can structured prompting reduce hallucination rates and improve accuracy?

Keywords: medical hallucinations, large language models, prompt engineering, Gemini, ChatGPT, transformer embeddings, RTHH

Introduction

Large language models can answer questions in many subjects, including medicine. While these models are helpful and sound confident, they sometimes produce information that is false or not supported by medical facts. These mistakes are called hallucinations. In a medical context, hallucinations can confuse users, spread misinformation, or cause harm.

This project focuses on improving the reliability of LLMs when answering medical questions. To study this, I developed RTHH - Robby the Hallucination Hunter which evaluates how two popular models, Gemini and ChatGPT, behave when the same medical questions are asked in two different ways: first with a simple prompt, and then with a structured prompt designed to reduce errors. By comparing both phases, the project shows how much impact prompt design alone can have on model accuracy.

Literature Review

Large language models like Gemini and ChatGPT are built on the Transformer architecture, which processes text using

attention mechanisms. This design is the foundation of most modern AI language systems. Even though the internal details of Gemini and ChatGPT are not visible to the user, knowing that they use transformers helps explain how they generate answers and why they sometimes make mistakes.

To measure how close a model's answer is to the correct medical answer, this project uses MiniLM-L6, a lightweight embedding model. MiniLM-L6 creates a vector for each answer, and cosine similarity is used to compare the model's answer to the ground truth answer. This allows the project to measure meaning, not just exact wording, which is helpful for evaluating medical responses.

Previous research shows that hallucinations are a well known issue in LLMs. The TruthfulQA benchmark found that models often give confident but false statements when asked factual questions. This problem becomes more serious in medicine. A study published in npj Digital Medicine showed that models sometimes add inaccurate medical details in clinical summaries, which can be unsafe.

One area of research that tries to reduce hallucinations is prompt engineering. The Chain-of-Verification paper showed that giving the model clear and structured instructions can reduce errors. This project uses that idea by creating structured prompts in Phase 2. These prompts help the models answer more carefully and avoid adding unsupported medical information.

Together, these findings support the need for a framework like RTHH and encourage testing whether structured prompts can reduce hallucinations in medical question answering tasks.

Research Approach

This project uses a two phase design to test how prompt style affects medical hallucinations in Gemini and ChatGPT.

The main dataset used is MedQuad, a medical question and answer dataset that provides expert approved answers. This dataset is used in both phases of the project. A smaller dataset from CuraiHealth was used only at the start of the project to test the pipeline, making sure the API calls and code were working correctly.

In Phase 1, each medical question was sent to Gemini and ChatGPT using a very simple prompt such as:

"Answer the following medical question."

This baseline prompt does not include any instructions about accuracy, safety, or uncertainty. The purpose of this phase is to observe the "natural" behavior of both models without guidance. Their answers were manually labeled as Correct, Incorrect, or Hallucinated, depending on whether the response matched the ground truth information from the dataset. A hallucination was defined as any added detail, symptom, treatment, or explanation that was not supported by the real answer.

To support the manual labels, the MiniLM-L6 embedding model was used to convert each answer into a vector and compute cosine similarity between the model's answer and the dataset's answer. Higher similarity scores generally matched correct

responses, while lower scores were often associated with hallucinations. This provided a consistent numerical way to compare Phase 1 and Phase 2.

In Phase 2, the same questions were asked again, but with structured prompts designed to reduce hallucinations. Four different prompt tones were tested:

Professional Doctor

Medical Specialist

Friendly / Respectful

Rude / Direct

These tones were chosen to explore how different styles of prompting affect the model's behavior. While all four tones worked on Gemini, only the rude / direct prompt produced reliable structured responses on ChatGPT. This happened because ChatGPT tends to give shorter and more focused answers when the tone is more direct, while Gemini performs better when given more detailed or professional context.

Even though the exact wording of each prompt is not included in this paper, all four tones shared the same core structure. They asked the model to use factual medical information, avoid guessing, and say when it was not sure. These structured prompts encouraged the models to be more careful and reduced the amount of unsupported information in their answers.

All responses from both phases were saved in CSV and JSONL files, and the results were summarized using three main diagrams: the baseline results, the ChatGPT prompt-tone results, and the Gemini prompt-tone results.

Results and Summary

This section presents the three main figures used to evaluate the project: the baseline test, the ChatGPT structured-prompt results, and the Gemini structured-prompt results.

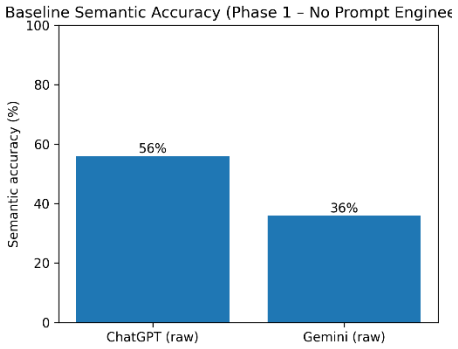

Figure 1: Baseline Model Results (Phase 1 - Simple Prompt)

Figure 1 shows how both models performed when answering medical questions with the simple baseline prompt. Both Gemini and ChatGPT produced a mixture of correct answers, incorrect answers, and hallucinations. Many hallucinations involved adding extra symptoms, risks, or medical explanations not found in the dataset's verified answers. The cosine similarity scores from the embedding model also tended to be lower in this phase.

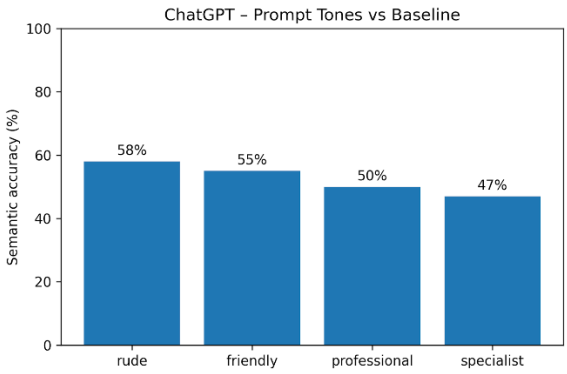

Figure 2: ChatGPT Structured Prompt Results (Phase 2)

Figure 2 shows that structured prompting did not always improve ChatGPT. Three of the four tones lowered accuracy and increased hallucinations. Only the "rude" tone improved ChatGPT's performance, producing shorter and more targeted answers with fewer mistakes.

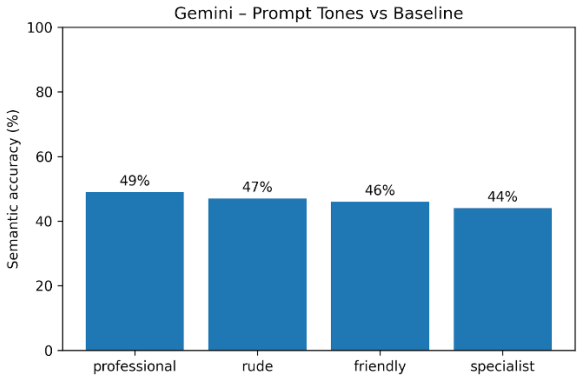

Figure 3: Gemini Structured Prompt Results (Phase 2)

Figure 3 shows Gemini's Phase 2 results. All four prompt tones improved performance compared to the baseline. Gemini gave more correct answers, produced fewer hallucinations, and scored higher on embedding similarity across every tone. This shows Gemini responds well to structured prompts no matter the style.

Overall, the diagrams show that structured prompting affects the two models differently. Gemini improved under all four tones, while ChatGPT improved only with the rude/direct prompt and dropped in accuracy with the others.

Conclusion

This project showed that large language models can change their medical accuracy depending on the way a question is asked. ChatGPT was the most accurate overall, but Gemini showed the biggest improvement when the prompt tone changed. Across both models, structured prompting influenced how often hallucinations appeared, and the embedding scores supported the manual labels. Even small changes in prompt style were enough to shift model behavior, proving that prompt design can make medical answers safer. Future work could test more models, including doctor reviewed scoring, or explore verification methods, but the results here show that prompt engineering is a practical first step toward reducing medical hallucinations.

References

Datasets

[1] Curai Health, "medical_questions_pairs." HuggingFace, 2023. [Online]. Available: https://huggingface.co/datasets/curaihealth/medical_questions_pairs

[2] K. Lia, "MedQuad – Medical Question Answering Dataset." HuggingFace, 2022. [Online]. Available: https://huggingface.co/datasets/keivalya/MedQuad-MedicalQnADataset

Transformers, Embeddings, and Prompt Engineering

[3] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019.

[5] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” EMNLP, 2019.

Hallucination & Medical Safety Literature

[6] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” arXiv preprint arXiv:2109.07958, 2021.

[7] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta, “A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation,” npj Digital Medicine, vol. 8, no. 1, article 274, May 2025.

[8] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, “Chain-of-Verification Reduces Hallucination in Large Language Models,” in Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, Aug. 2024.