RTHH: Robby The Hallucination Hunter

Evaluating Medical Hallucinations in Large Language Models

Arturo G. Valle

December 3, 2025

Why This Research Matters

- LLMs are widely used for medical questions.
- Many people rely on them due to cost, convenience, or lack of access to a doctor.
- Hallucinations give very confident but incorrect answers.
- Dangerous when people use LLMs instead of doctors.

The Problem

- When an LLM invents or produces false medical information. It comes with risks like Wrong self-treatment, Unsafe recommendations, and Delayed care.
- Existing work that I was able to find is not medical focused.

Project Contributions

- Built a medical hallucination evaluation pipeline.
- Compared two LLMs: ChatGPT vs Gemini.
- Tested four prompt tones: professional, specialist, friendly, and rude.
- Produced baseline (Raw Test) and prompt-engineered results.

Methodology Overview

- **Dataset:** 150 medical questions from MedQuad- MedicalQnADataset https://huggingface.co/datasets/keivalya/MedQuad-MedicalQnADataset .
- **Models Tested:** ChatGPT (gpt-4o-mini)and Gemini (gemini-2.5-flash)
- **Evaluation Metric:** Semantic accuracy (0–100%), Hallucinations were incorrect when the response was off-topic, fabricated, or had unsafe answers.

Methodology Overview

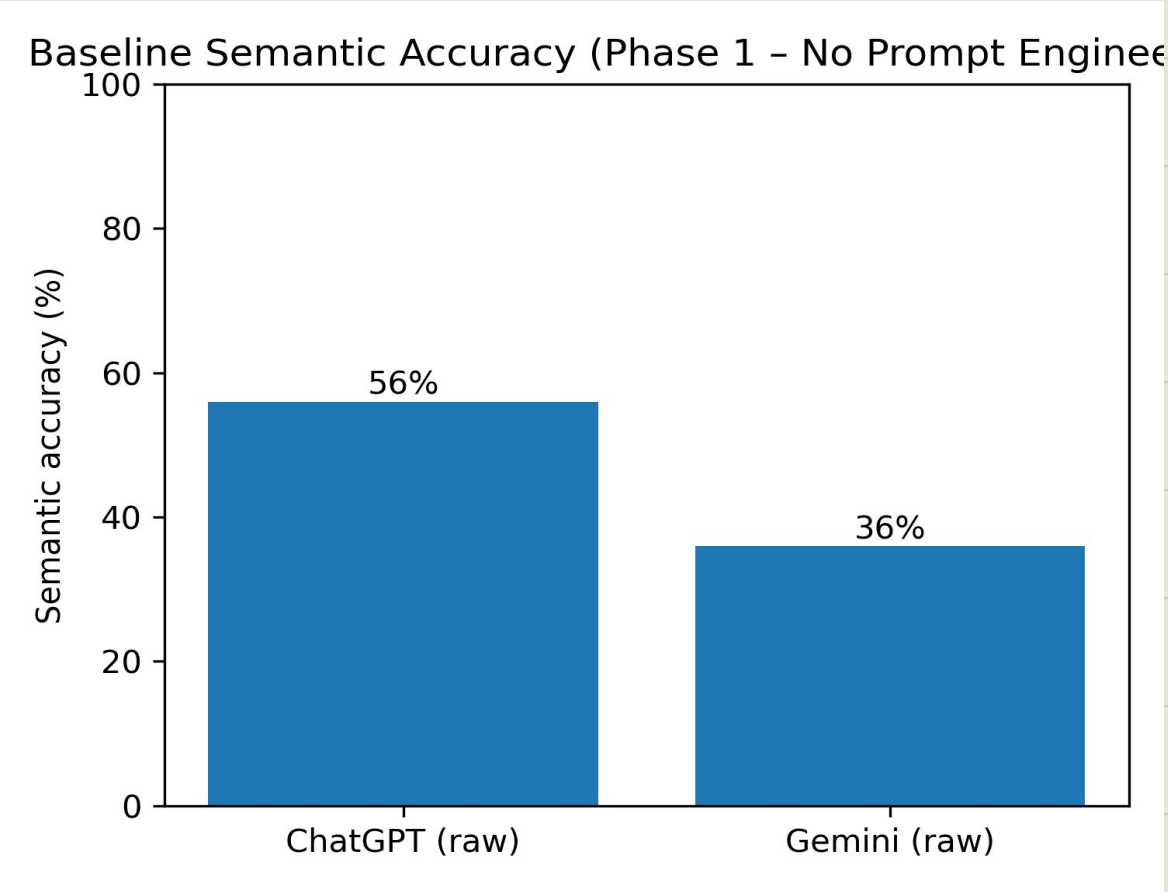
Embeddings:

- Used MiniLM-L6, a transformer embedding model, to convert answers into vectors.
- Used cosine similarity to compare model answers to reference answers.
- Chosen because it is:
  - ➔ Optimized for semantic similarity
  - ➔ Lightweight and fast
  - ➔ Most importantly, stable across different prompt tones, ensuring consistent evaluation
- Project had 2 phases: Phase 1 - Raw prompts, and Phase 2 - Prompt tones (Prompt Engineering)

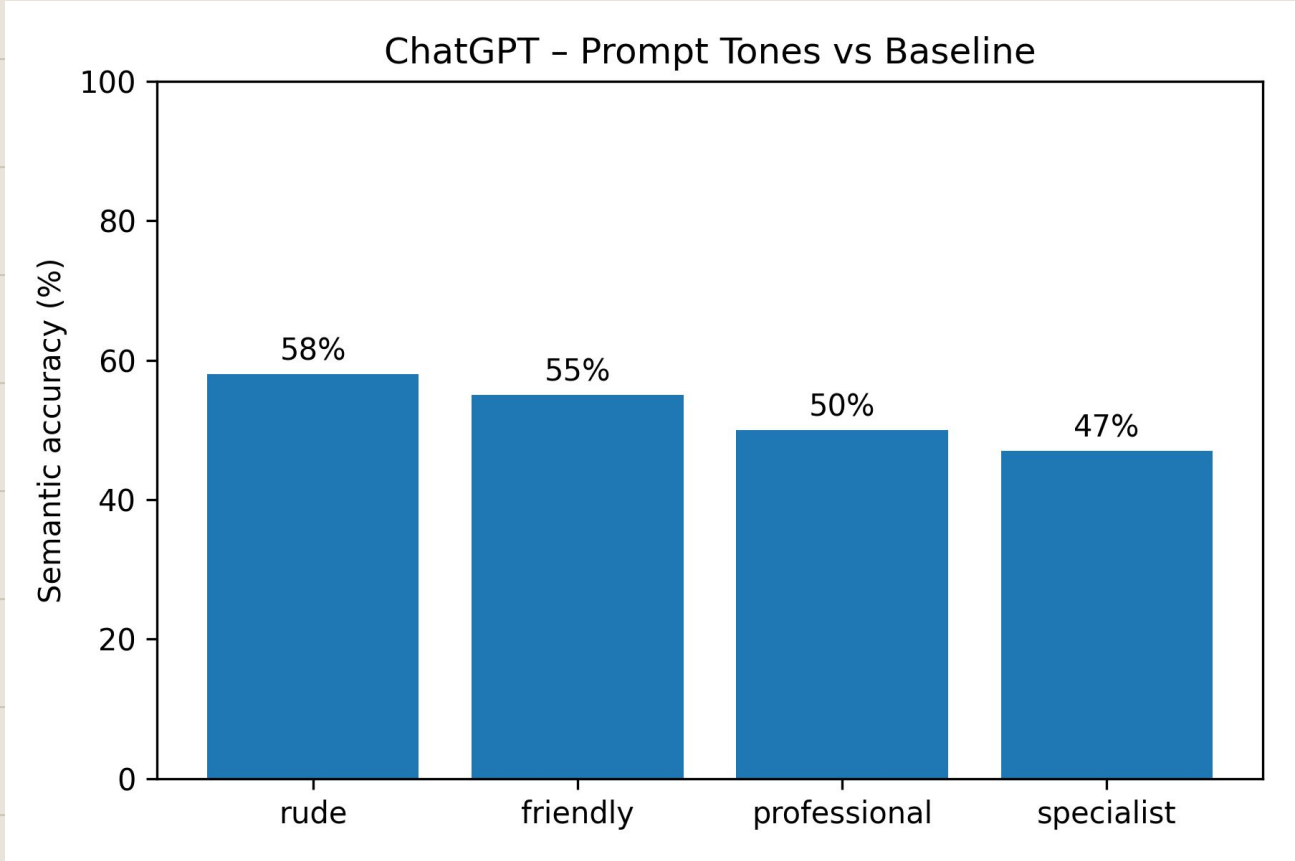
Experimental Setup

- Same 150 medical questions used for all conditions to ensure a fair comparison.
- Four prompt tones tested because prior prompt engineering research shows that tone, framing, and phrasing can influence model behavior.
- Same scoring pipeline (MiniLM embeddings + cosine similarity) used for all outputs, that ensured a consistent measurement.

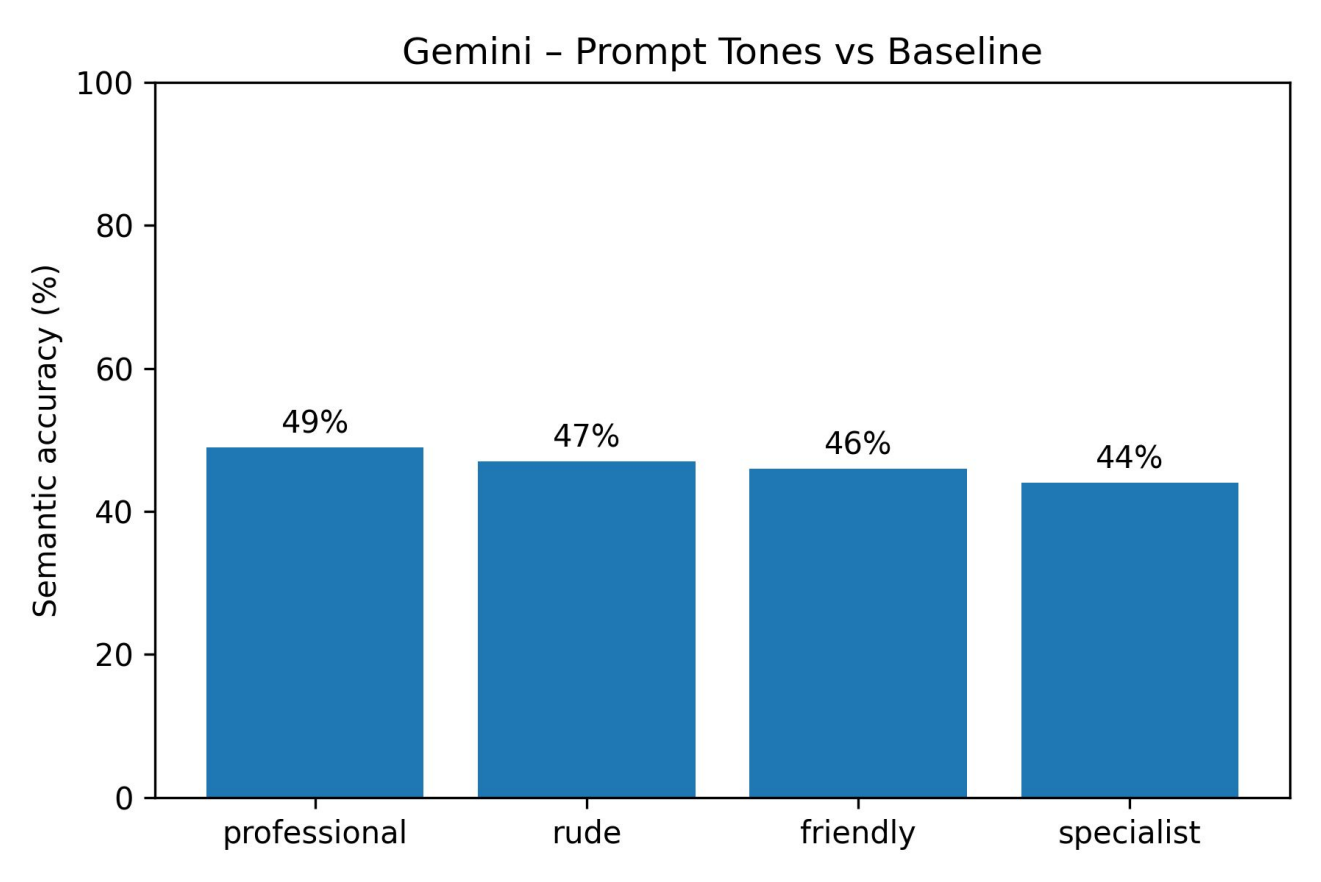
Baseline Results



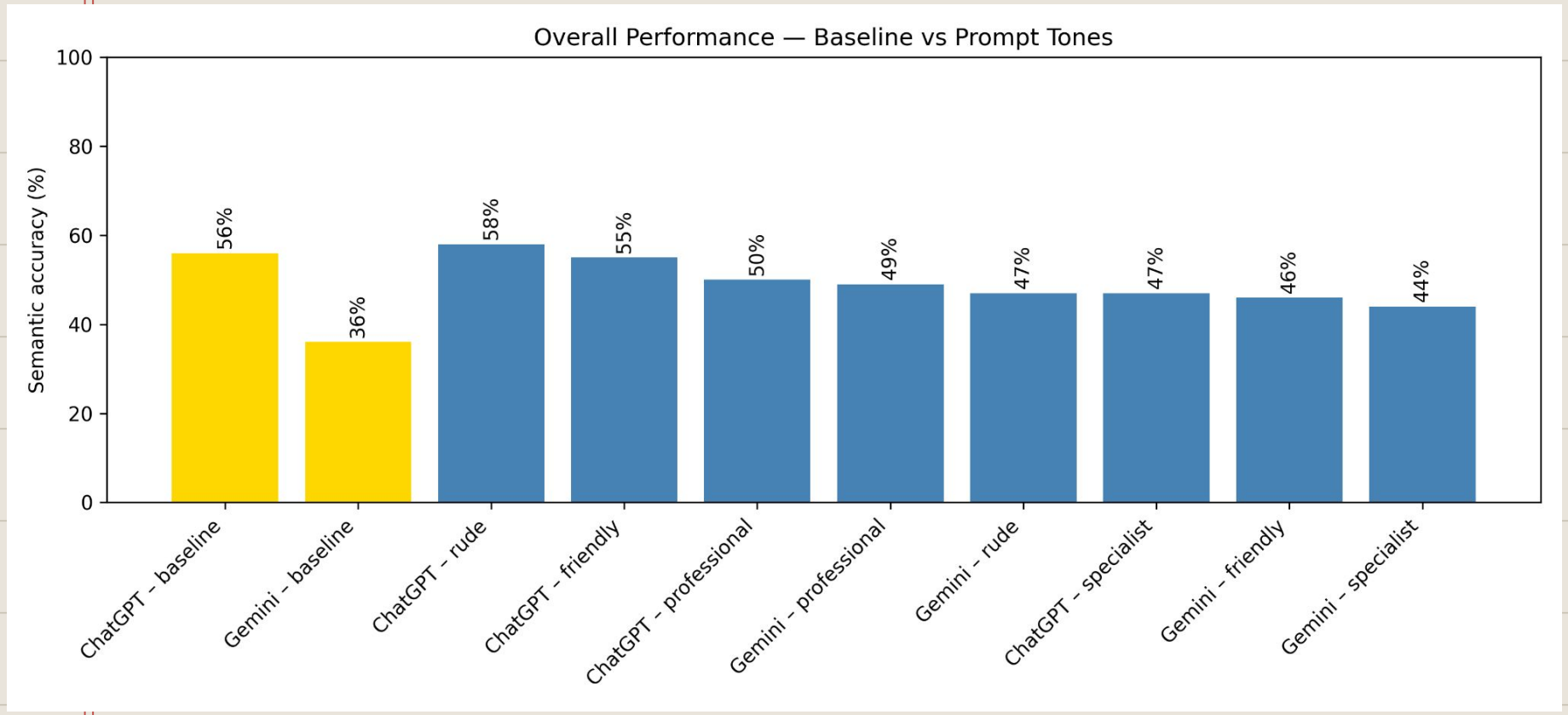
ChatGPT Results (Prompt Tones)



Gemini Results (Prompt Tones)



Overall Performance



Comparison With Existing Approaches

TruthfulQA (2021)

- A benchmark that tests whether AI gives truthful answers.
- Covers many topics, but not focused on medicine.

Chain-of-Verification (2024)

- A method where the AI checks its own work to reduce mistakes.
- Designed to fix hallucinations.

How Robby Is Different

- Measures medical hallucinations, not general truthfulness.
- Observes the model's behavior
- Studies how prompt tone affects accuracy.

Key Findings

- ChatGPT was the most accurate overall.
- Gemini improved the most when the tone changed.
- Prompt tone does affect hallucinations.
- Embeddings (MiniLM-L6) gave a consistent, meaning-based scoring.

Future Work

- Expand beyond 150 medical questions.
- Test more models (Copilot, Grok, etc).
- Add hallucination severity levels (minor vs dangerous).
- Experiment with mitigation methods like Chain of Verification (CoVe).

Thank You, Any Questions?