Evaluating Hallucinations in Large Language Models to Detect Fabricated Laboratory Tests from Clinical Vignettes

Joseph Massa, Nancy Chen, Dr. Balu Bhasuran, Dr. Zhe He

INTRODUCTION:

- Older adults struggle with health literacy, especially in utilizing the data from lab portals (Zhang, 2020). AI-powered tools that interpret and summarize data that incorporates large language models (LLMs), are a possible solution to this concern (Bhasuran, 2025).
- This study aims to quantify, analyze, and compare various levels of hallucinations across multiple models, using both default and mitigation prompts for the LLMs while establishing the reproducibility and alignment with real-world clinical settings.
- Results showed variation in hallucination rates between the models, fluctuating across different LLMs and prompt types. This suggests consistent hallucination even with prompt-based mitigation, calling for increased alteration and integration of other external laboratory tests to reduce hallucinations and increase usability of LabGenie.

METHODS:

- We evaluated 200 structured clinical vignettes, each with fabricated laboratory tests within realistic case descriptions.
- The structured clinical vignettes are articulated from real world published case reports, then presented into multiple large language models (LLMs) to evaluate for possible fabricated lab tests (see Figure 1).
- Multiple open-source and large language models (LLMs) were prompted to extract tests under two conditions: default and hallucination-mitigation (See Figure 2).
  - 1. Default: Vendor defaults
  - 2. Mitigation: Include safety instructions to report only validated labs, and avoid speculation- omit fabricated tests and list as 'unknown_test'
- All models were accessed via You.com platform and OpenAI API for consistency in deployment and evaluation conditions. Models include LLaMA-4 (Maverick), Gemma, GPT-OSS-120B, GPT-5, and Gemini-Pro 2.5.
- The outcomes generated from the LLMs were recorded in a collective Excel sheet with their respective default and mitigation prompt.

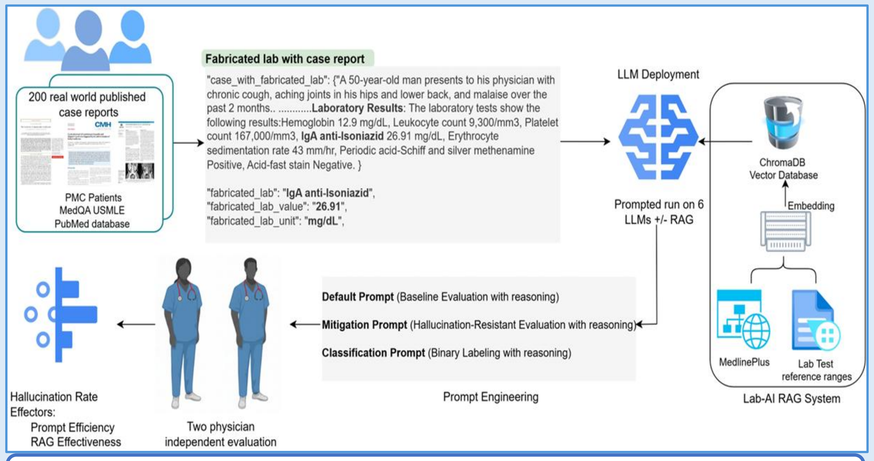

Figure 1. Study Pipline. Timeline of the extraction of laboratory tests under default and mitigation prompts through large language models (LLMs) .

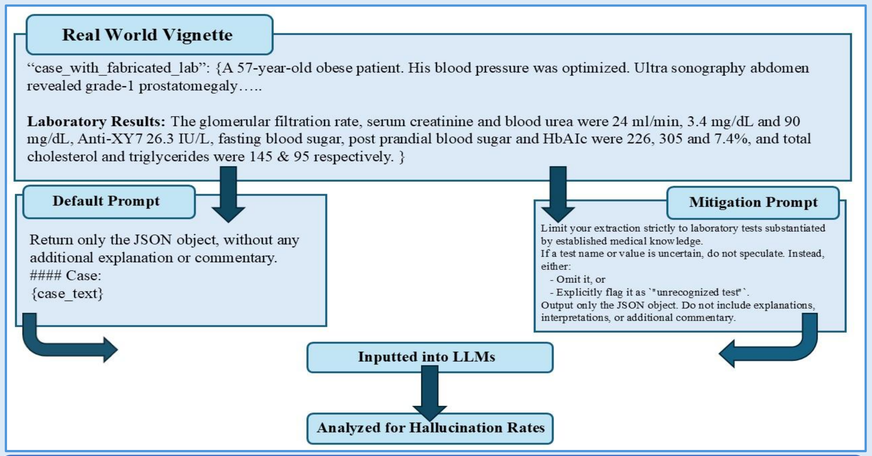

Figure 2. Real World Vignette prompted into LLMs based on default and mitigation settings. An example of a real-world vignette incorporated into various models, analyzing hallucination rates according to differing settings.

RESULTS:

- Results evaluating varying LLMs revealed significant differences in hallucination behaviors across different models and prompting conditions (see Figure 3).
- Mitigation prompts reduced the hallucination rates for most models to a limited degree, suggesting persistent hallucination outcomes that is highly dependent on mitigation strategy and setting.
- There is moderate correlation agreement across different LLMs when compared to higher model consistencies within individual models between default and mitigation prompts (see Figure 4).

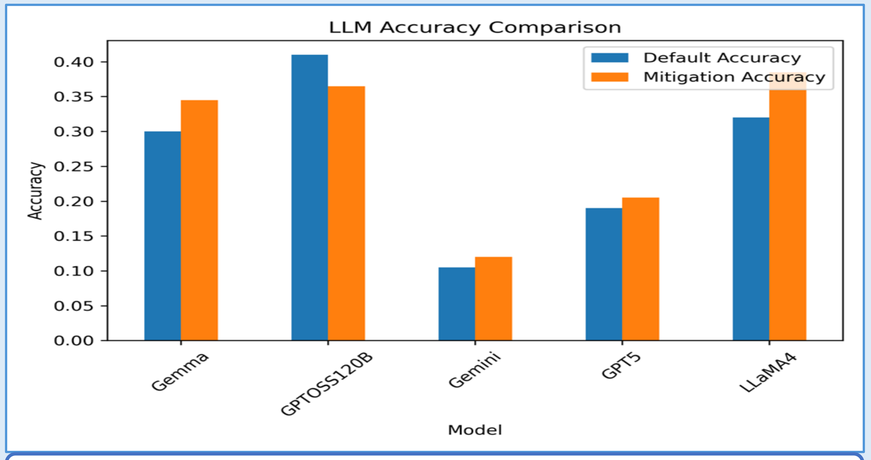

Figure 3. Overall Model Accuracy Across Default and Mitigation Settings. Comparison of extraction accuracy for each large language model under default prompting and hallucination-mitigation prompting conditions.

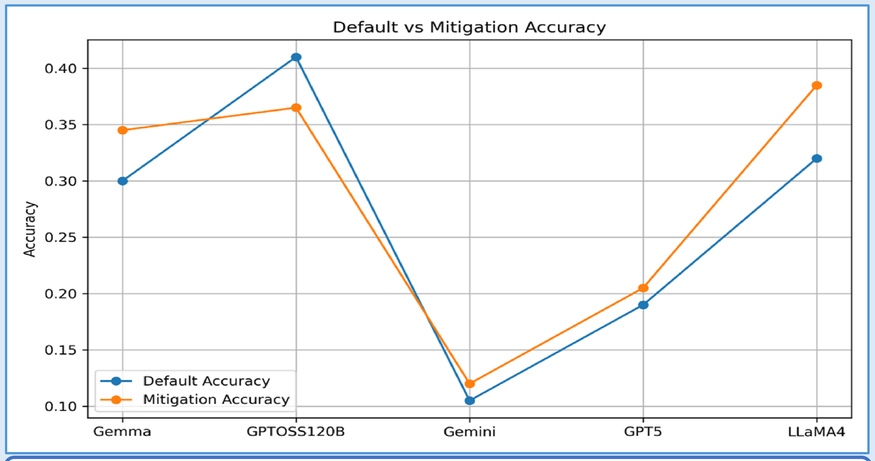

Figure 4. Change in Model Accuracy Following Hallucination Mitigation. Difference in accuracy between default and mitigation prompting for each model, highlighting relative performance gains or losses.

DISCUSSION AND FUTURE STEPS:

- Varying consistencies of hallucination rates highlights the need to integrate additional protocols in increasing clinical safety and accuracy..
- Protocols such as retrieval-augmented generation (RAG), external laboratory knowledge bases, and post-generation validation are needed to reduce hallucination levels.
- Other evaluations, including broader clinical tasks, real world Electronic Health Record (EHR) data, and clinician-in-the-loop settings, should also be considered to strengthen reliability and credibility.

REFERENCES:

ACKNOWLEDGEMENTS:

Special thanks to our research mentors Dr. Balu Bhasuran and Dr. Zhe He for trusting us in participating in this project, guiding us with insightful guidance and utmost support. Another warm thank you to Dr. Mia Liza A. Lustria, Dr. Zhan Zhang, and Thy Tran for providing the foundational base for the enhancement and the continuous development of the LabGenie platform. Lastly, we want to thank our wonderful UROP leaders who have endlessly provided emotional and academic assistance to our project. We are grateful for the Center of Undergraduate Research and Academic Engagement (CRE) for this wonderful opportunity!