# Informatics Supporting Patients' Understanding of Lab Results: Identifying Patients' Questions about Lab Results

Maggie Awad, Yash Alva and Dr. Zhe He

## INTRODUCTION

Large language models (LLMs) such as ChatGPT have opened a promising avenue for patients to get their questions answered. We aim to assess the feasibility of using LLMs to generate relevant, accurate, helpful, and unharmful responses to lab test-related questions asked by patients and to identify potential issues that can be mitigated with augmentation approaches. We believe that with the advancements of Ai and LLMs it has opened up many new possibilities when it comes to interpreting and understanding health lab reports. However, even though generative AI models such as ChatGPT can answer questions, about lab test results, they may also generate answers with inaccurate information or hallucinations results and are often confusing and hard to understand. We believe by utilizing the advantages of AI and LLMs, we will be able to create an app that will help patients (especially elderly patients) better understand their health lab reports

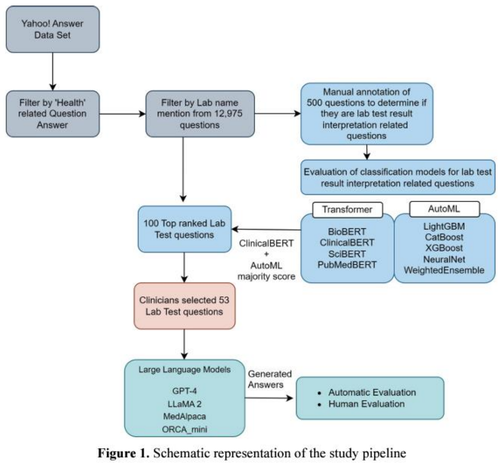


**Figure 1.** Schematic representation of the study pipeline

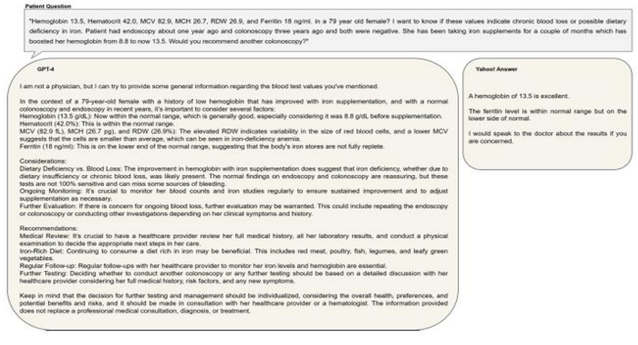


**Figure 2**. Responses from GPT-4 and a human for an example lab result interpretation question from Yahoo! Answers.

## METHODS

**Overview:** We first collected lab test results related question and answer data from Yahoo! Answers and selected 53 QA pairs for this study. We generated responses to the 53 questions from four LLMs including GPT-4, Meta LLaMA 2, MedAlpaca, and ORCA_mini. We first assessed the similarity of their answers using standard QA similarity-based evaluation metrics including ROUGE, BLEU, METEOR, BERTScore. Finally, we performed a manual evaluation with medical experts for all the responses to seven selected questions on the same four aspects.

**Categorizing Data:** We entered lab test result questions from Yahoo! answers into 4 different LLMs and asked the LLM to figure out if the given questions were lab related or not.

**Interpreting Data:** After categorizing the data, we then checked to see if the LLMs could provide the correct answer to the medical questions and give an accurate and helpful response.

**Analyzing Data:** To see which LLM performed the best we used an evaluator to see which of the LLMs performed better over the categories of correctness, relevance, helpfulness, and safety. In order to ensure that the LLMs responses were accurate, we began a new chat each time there was information provided to a LLM to eliminate biases.

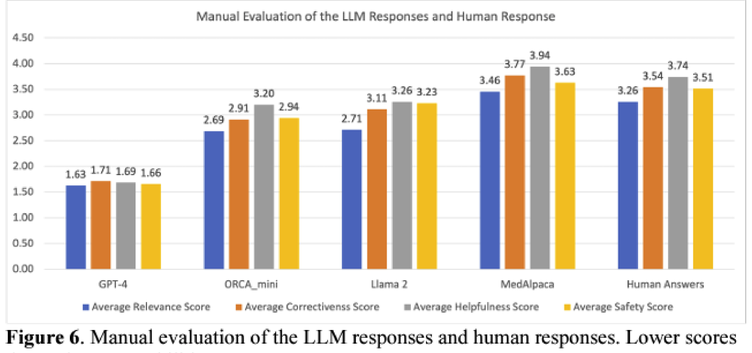


**Figure 6**. Manual evaluation of the LLM responses and human responses. Lower scores denote better capabilities.

## RESULTS

In our initial study we saw that GPT-4 performed the best out of the 4 LLMs we tested achieving better scores in relevance, correctness, helpfulness, and safety. Even though it performed the best we noticed that there was still occasional errors in one of those 4 categories in each LLM. Our Results are currently still in the preliminary stages, and we are still analyzing and selecting different case studies and performing literature reviews to better understand how AI responses can be used to interpret lab report data. As for the LabGenie app we are currently working on getting a developer so we can go further with designing and coding an app interface that will interpret health lab reports and make it easier for patient to understand their lab reports.

## CONCLUSION

Our current results show that Ai tools and LLMs are able to provide accurate and helpful information when interpreting lab report data. And in this study, we saw that Chat GPTs GPT-4 performed the best out of the 4 LLMs we tested. However, there's is still a gap in consistency as we saw there were instances where these LLMs were not able to provide accurate or helpful information that was relevant to the patients scenario. By analyzing more cases and doing more literature review we will have a better understanding of the full capabilities of Ai and LLMs in terms oof interpreting lab report data and we will be able to use this information to develop an app that will help patients better understand their health lab reports.

## REFERENCES

* He, Z., Bhasuran, B., Jin, Q., Tian, S., Hanna, K., Shavor, C., Arguello, L. G., Murray, P., & Lu, Z. (2024, January 23). *Quality of answers of Generative Large language models vs peer patients for interpreting lab test results for Lay Patients: Evaluation Study*. arXiv.org. https://arxiv.org/abs/2402.01693