Evaluating Large Language Models for Accurate Lab Test Question Interpretation

Caroline Bennett, Angelique Deville, Hailey Thompson, Dr. Zhe He & Dr. Balu Bhasuran

College of Communication & Information, Florida State University

Introduction

The LabGenie project seeks to empower patients of low health literacy, particularly the elderly, to ask contextualized questions and make informed health decisions. Through the integration of retrieval-augmented generation (RAG)-based large language models (LLMs) and a user-friendly platform, LabGenie will generate answers based on accurate health information. In order to develop LabGenie, we must conduct preliminary research on the viability of using LLMs for clinical purposes. Our research currently involves two studies: 1) evaluating the strengths and weaknesses of different LLMs in answering lab result questions and 2) calculating the accuracies of LLMs when developing differential diagnoses. We predict that large language models (LLMs) such as ChatGPT have opened a promising avenue for patients to get their questions answered.
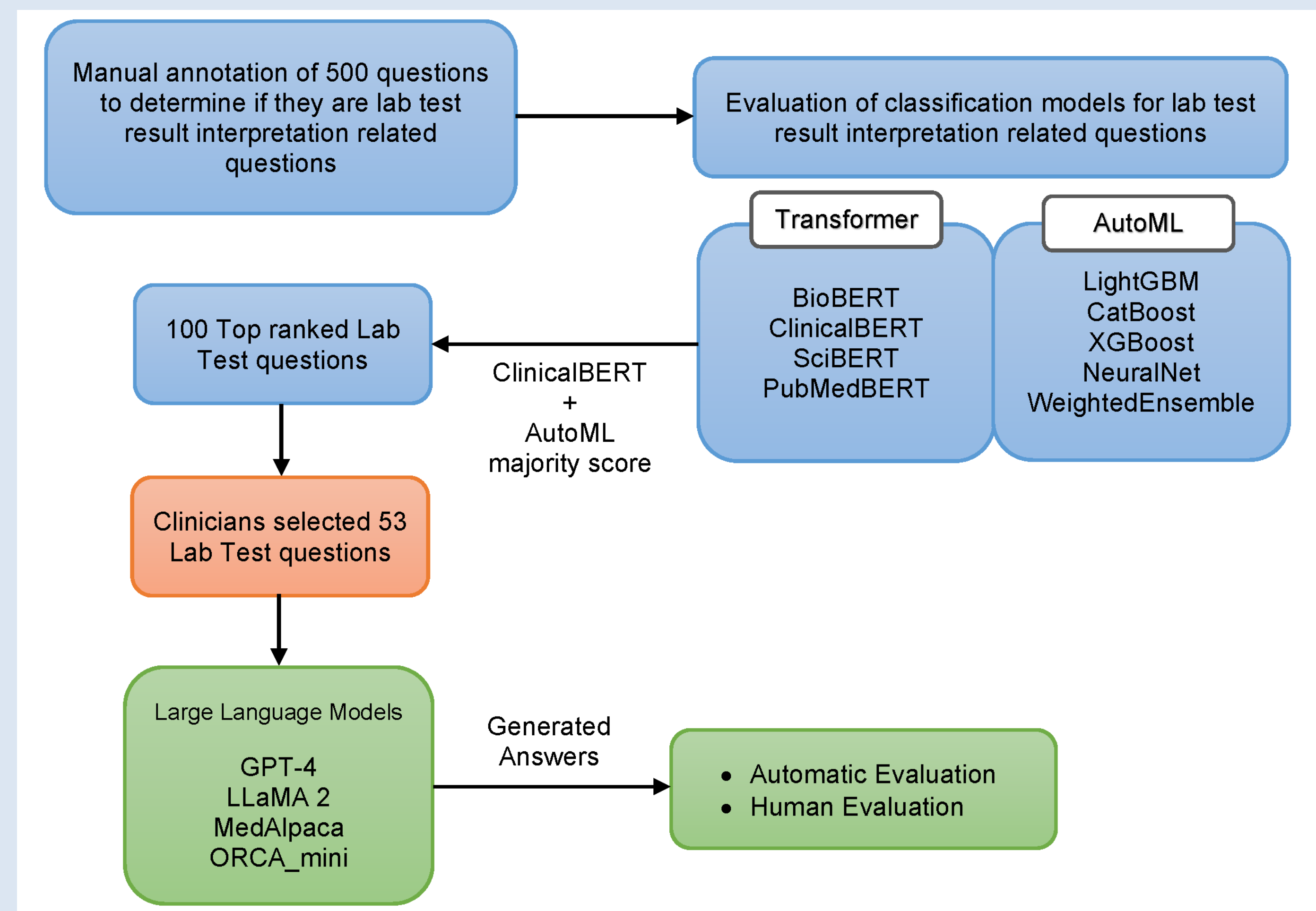
Project 1 – Methods & Results

Figure 1. Schematic representation of methods for Project 1.

Figure 2. Example lab test interpretation from GPT-4 and a human for Project 1.
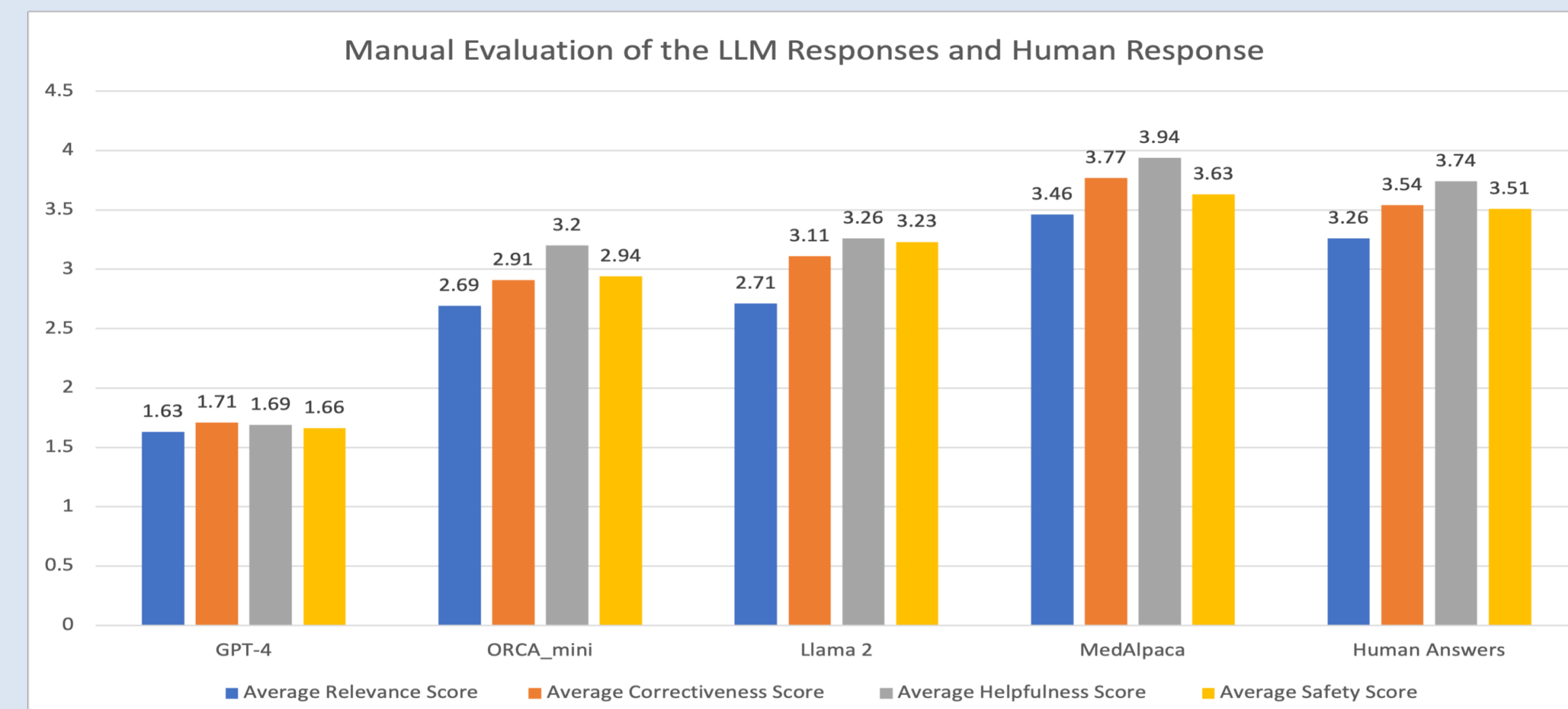
Project 1 – Methods & Results (cont.)

Figure 3. Schematic representation of methods for Project 2.
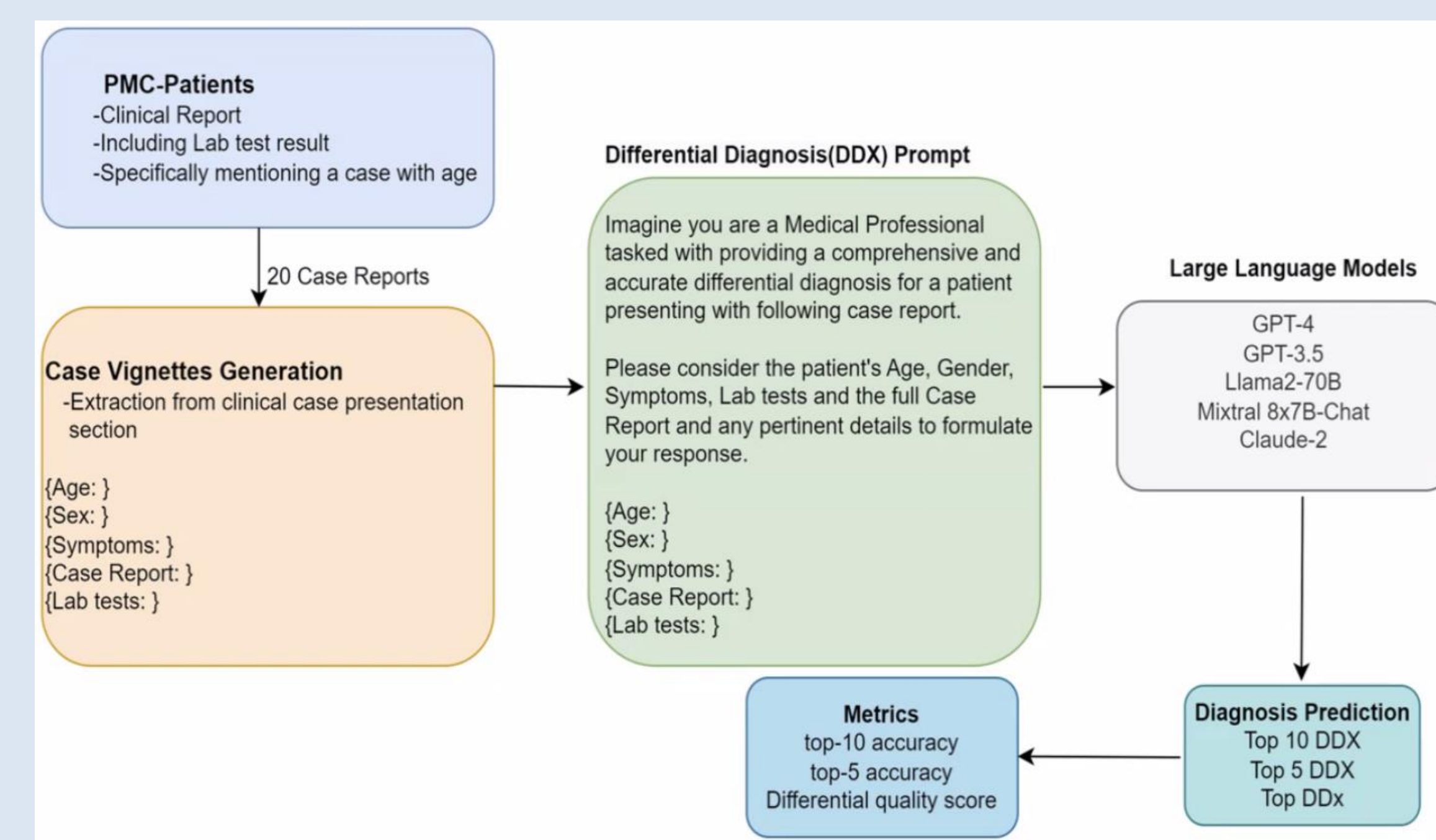
Project 2 – Methods & Results

Figure 4. Results of Project 1. Lower scores denote better capabilities.

Figure 5. LLMs accuracies in diagnosis prediction for Project 2.

Discussion

Principal Results:

1) Based on the results of our study, GPT-4 outperformed other similar LLMs, such as Llama-2, MedAlpaca and ORCA_mini, according to both automated evaluation and manual evaluation results. For the best results, it is important to provide follow-up questions for additional information, to understand LLMs should be used for clarification instead of diagnosis, and to recognize that the validity of answers are hard to determine without the review of medical professionals.

2) With lab test information provided, the results for the LLMs were fairly variable. LLMs that had a low accuracy for one requested diagnosis, such as Claude2 and Llama-2, saw an increase in accuracy as more diagnoses were requested. Meanwhile, LLMs with a higher initial accuracy either stayed the same or decreased as more diagnoses were requested. When lab test information was not requested, there is an evident peak in accuracy across all the LLMs when 5 LLMs were requested. Accuracy was notably higher for each LLM with lab tests than without lab tests.

Future Directions:

1) Encourage thorough inquiries between the LLM and the user, have medical professionals review LLM responses, experiment with prompt engineering, provide standardized reference ranges, evaluate confidence levels of the sentences, and develop gold-standard answers.

2) Increase sample size of case reports and ask the LLMs how the results of specific lab tests influences their decision when forming diagnosis.

Conclusions

1) GPT-4 provided the most relevant, correct, helpful, and safe responses compared to the other LLMs. However, while GPT-4 may be feasible for lab test interpretation, there are still risks of hallucinations, which need to be addressed before being used for diagnostic purposes.

2) Each of the LLMs had variable results when requested different numbers of differential diagnoses as well as whether lab test results were given or excluded. Without further research, we conclude that there are still risks when utilizing LLMs for diagnosis, especially without lab test results.

References

He, Z., Bhasuran, B., Jin, Q., Tian, S., Hanna, K., Shavor, C., Arguello, L. G., Murray, P., & Lu, Z. (2024). Quality of Answers of Generative Large Language Models vs Peer Patients for Interpreting Lab Test Results for Lay Patients: Evaluation Study. arXiv preprint arXiv:2402.01693.

Acknowledgements