



# 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

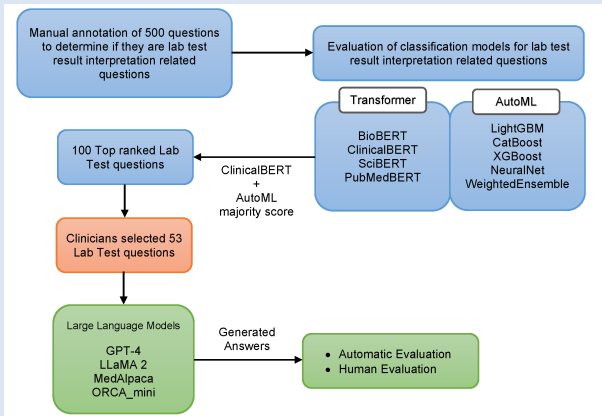


Figure 1. Schematic representation of methods for Project 1.

**Human Answer**

"Hemoglobin 13.5, Hematocrit 42.0, MCV 82.9, MCH 26.7, RDW 26.9, and Ferritin 18 ng/ml. In a 79 year old female? I want to know if these values indicate chronic blood loss or possible dietary deficiency in iron. Patient had endoscopy about one year ago and colonoscopy three years ago and both were negative. She has been taking iron supplements for a couple of months which has boosted her hemoglobin from 8.8 to now 13.5. Would you recommend another colonoscopy?"

**GPT-4**

I am not a physician, but I can try to provide some general information regarding the blood test values you've mentioned.

In the context of a 79-year-old female with a history of low hemoglobin that has improved with iron supplementation, and with a normal colonoscopy and endoscopy in recent years, it's important to consider several factors:

**Hemoglobin (13.5 g/dL):** This value is within the normal range, which is generally good, especially considering it was 8.8 g/dL before supplementation.

**Hematocrit (42.0%):** This is also within the normal range.

**MCV (82.9 fL), MCH (26.7 pg), and RDW (26.9%):** The elevated RDW indicates variability in the size of red blood cells, and a lower MCV suggests that the cells are smaller than average, which can be seen in iron-deficiency anemia.

**Ferritin (18 ng/ml):** This is on the lower end of the normal range, suggesting that the body's iron stores are not fully replete.

**Considerations:**

**Dietary Deficiency vs. Blood Loss:** The improvement in hemoglobin with iron supplementation does suggest that iron deficiency, whether due to dietary insufficiency or chronic blood loss, was likely present. The normal findings on endoscopy and colonoscopy are reassuring, but these tests are not 100% sensitive and can miss some sources of bleeding.

**Ongoing Monitoring:** It's crucial to monitor her blood counts and iron studies regularly to ensure sustained improvement and to adjust supplementation as necessary.

**Further Evaluation:** If there is concern for ongoing blood loss, further evaluation may be warranted. This could include repeating the endoscopy or colonoscopy or conducting other investigations depending on her clinical symptoms and history.

**Recommendations:**

**Medical Review:** It's crucial to have a healthcare provider review her full medical history, all her laboratory results, and conduct a physical examination to decide the appropriate next steps in her care.

**Iron-Rich Diet:** Continuing to consume a diet rich in iron may be beneficial. This includes red meat, poultry, fish, legumes, and leafy green vegetables.

**Regular Follow-up:** Regular follow-ups with her healthcare provider to monitor her iron levels and hemoglobin are essential.

**Further Testing:** Deciding whether to conduct another colonoscopy or any further testing should be based on a detailed discussion with her healthcare provider considering her full medical history, risk factors, and any new symptoms.

**Keep in mind:** that the decision for further testing and management should be individualized, considering the overall health, preferences, and potential benefits and risks, and it should be made in consultation with her healthcare provider or a hematologist. The information provided does not replace a professional medical consultation, diagnosis, or treatment.

Figure 2. Example lab test interpretation from GPT-4 and a human for Project 1.

## Project 1 – Methods & Results (cont.)

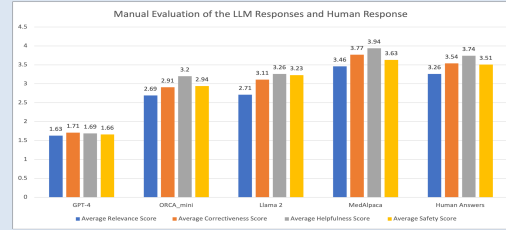


Figure 3. Schematic representation of methods for Project 2.

## Project 2 – Methods & Results

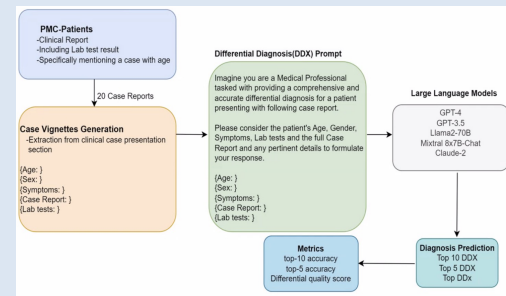


Figure 4. Results of Project 1. Lower scores denote better capabilities.



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

## Discussion

Principal Results:

- 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.
- 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:

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

## Conclusions

- 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.
- 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

We would like to thank Dr. He and Dr. Bhasuran for the opportunity to help them on this research project and for being supportive throughout the learning process. We would also like to thank all the researchers involved in the eHealth Lab for assisting with this project.



eHealth Lab