2025 ANNUAL SYMPOSIUM
WHERE COLLABORATION MEETS HEALTHCARE TRANSFORMATION
#AMIA2025

Preliminary Analysis of the Impact of Lab Results on Large Language Model Generated Differential Diagnoses

Balu Bhasuran[1], Qiao Jin[2], Yuzhang Xie[3], Carl Yang[3], Karim Hanna[4], Jennifer Costa[4], Cindy Shavor[4], Wenshan Han[5], Zhiyong Lu[2], Zhe He[1,*]

[1]Florida State University, Tallahassee, FL; [2]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD; [3]Emory University, Atlanta, GA; [4]Morsani College of Medicine, University of South Florida, Tampa, FL; [5]Department of Population and Community Health, University of North Texas Health Science Center, Fort Worth, TX

Introduction

Differential diagnosis is crucial for medicine as it helps healthcare providers systematically distinguish between conditions that share similar symptoms[1,2].

This study assesses the impact of lab test results on differential diagnoses (DDx) made by large language models (LLMs).

Clinical vignettes from 50 case reports from PubMed Central were created incorporating patient demographics, symptoms, and lab results.

Five LLMs—**GPT-4, GPT-3.5, Llama-2-70b, Claude-2, and Mixtral-8x7B**—were tested to generate Top 10, Top 5, and Top 1 DDx with and without lab data. A comprehensive evaluation involving GPT-4, a knowledge graph, and clinicians was conducted.

Methods

Clinical case reports for this assessment were obtained from the PMC-Patients dataset. From 50 selected case reports, we manually generated clinical vignettes that included details such as patient age, gender, symptoms, laboratory test results, and other relevant information, allowing the models to generate differential diagnosis responses.

A specific prompt was designed to instruct the models to consider all relevant details and provide differential diagnoses, including Top 1, Top 5, and Top 10 DDx lists. Model predictions were reviewed by clinicians and automatically evaluated using a knowledge graph and GPT-4, utilizing exact match, relevance, and incorrect predictions.

The diagnostic accuracy was evaluated using exact and lenient accuracy metrics for Top 1, Top 5, and Top 10 differential diagnoses (DDx), derived from clinical vignettes with and without laboratory test data.

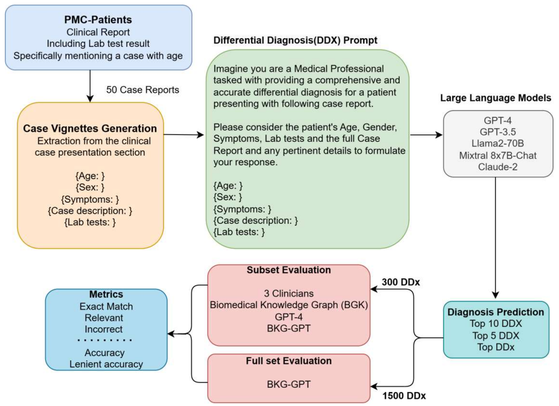

Figure 1. Study pipeline for evaluating large language models (LLMs) in differential diagnosis (DDX) generation.

Case Reports

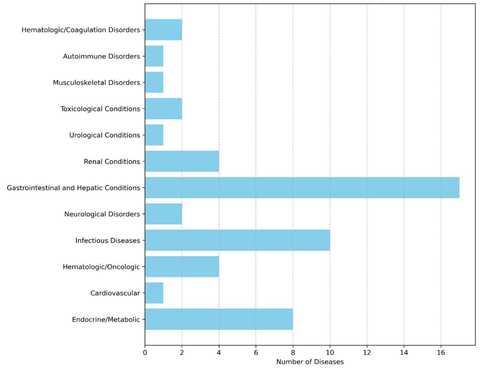

Figure 2. Distribution of diseases across medical categories in differential diagnosis evaluation.

Results

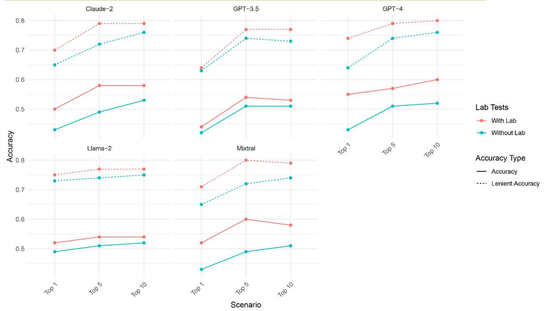

Figure 4. Effect of Lab Tests on Accuracy and Lenient Accuracy Across Scenarios.

GPT-4 consistently showed the highest performance in generating differential diagnoses across various scenarios, particularly excelling in lenient accuracy. Mixtral followed closely, especially in the Top 5 and Top 10 lists, while GPT-3.5 stood out for its high lenient accuracy in the Top 5 diagnoses. Claude-2 slightly outperformed LLaMa-2, particularly in exact match rates and lenient accuracy across different differential diagnosis lists.

Table 2. Top-k prediction accuracy comparison from Paired T-test and Holm's method evaluating the impact of lab test data

Top-k Prediction Comparison	P-value	Adjusted P-value
Top 1 with lab test data vs without lab test data	0.023	0.0483
Top 5 with lab test data vs without lab test data	0.016	0.0483
Top 10 with lab test data vs without lab test data	0.018	0.0483

All adjusted p-values using Holm's method are below 0.05, indicating statistically significant differences between models with and without lab data, even after correction for multiple comparisons.

Automatic Evaluation

"Please assess the accuracy of our predicted diagnosis in comparison to the true diagnosis.
Consider whether the prediction matches the true diagnosis directly, represents a variant or specific form of the true diagnosis, or is broadly correct but not exact.
For the given diagnoses: Predicted diagnosis: [ ],True diagnosis: [ ]
Evaluate the match according to the following criteria:
Exact Match: The predicted diagnosis is exactly the same as the true diagnosis
Relevant: The predicted diagnosis is a variant, form, closely related or refer to the same condition with slight variations in wording or captures the broad category or concept of the true diagnosis or broadly aligns with the true disease category but differs in specifics
Incorrect: The predicted diagnosis does not accurately reflect the true diagnosis
Please select the most appropriate option: [Exact Match/Relevant/Incorrect]
Provide your evaluation in the following JSON format:\n\n{\"evaluation\": \"Choose from: Exact Match, Relevant, Incorrect\",\"predicted_diagnosis\": \"[Your Predicted Diagnosis Here]\",\"true_diagnosis\": \"[Actual Diagnosis Here]\"}

For the given diagnoses:Predicted diagnosis: [Diabetic Nephropathy],
True diagnosis: [diabetic nephropathy with near-nephrotic range proteinuria]

GPT-4 Prediction
{\"evaluation\": \"**Relevant**\",\"predicted_diagnosis\": \"Diabetic Nephropathy\",\"true_diagnosis\": \"diabetic nephropathy with near-nephrotic range proteinuria\"}

Figure 3: Example of automatic evaluation differential diagnosis from LLMs using GPT-4

Discussion

This study evaluated the impact of lab test results on the accuracy of differential diagnoses using five LLMs with published clinical case reports. GPT-4 achieved the highest performance, with Top 1 accuracy of 55% (95% CI 0.41–0.69) and Top 10 accuracy of 60% (0.46–0.74) when incorporating lab data with lenient accuracy reaching 79% (0.68–0.90). Holm-adjusted p-values were all below 0.05, confirming statistically significant improvements with lab data with GPT-4 and Mixtral excelling, though exact match rates were low.

The PubMed search highlights the rarity of the majority of the diagnoses, as 70% of them are reported in fewer than 100 articles. Since these are such rare conditions, LLMs must possess specific knowledge of these diseases to make accurate diagnosis predictions.

Table 3. Disease Incidence Distribution Based on PubMed Literature Review for 50 Case Reports

Disease incidence range	Number of cases
1-10	22
11-100	13
101-1000	10
1001-10000	1
>10000	4

Clinician Evaluation

Table 1: Comparative Accuracy and Mismatch Analysis of 300 predictions

Evaluation	Agreement	Disagreement	Alignment Percentage (%)	Variance Percentage (%)
GPT-4 vs Clinician (First Scenario)				
Claude	45	15	75.00	25.00
GPT-3.5	43	17	71.67	28.33
GPT-4	44	16	73.33	26.67
LLaMa2	40	20	66.67	33.33
Mixtral	44	16	73.33	26.67
		Average	72	28
GPT-4 vs BKG (Second Scenario)				
Claude	39	21	65.00	35.00
GPT-3.5	52	8	86.67	13.33
GPT-4	47	13	78.33	21.67
LLaMa2	34	26	56.67	43.33
Mixtral	41	19	68.33	31.67
		Average	71	29
Clinician vs BKG (Third Scenario)				
Claude	48	12	80.00	20.00
GPT-3.5	49	11	81.67	18.33
GPT-4	55	5	91.67	8.33
LLaMa2	44	16	73.33	26.67
Mixtral	51	9	85.00	15.00
		Average	82.33	17.66
Clinician vs BKG-GPT(Fourth Scenario)				
Claude	50	10	83.33	16.67
GPT-3.5	52	8	86.67	13.33
GPT-4	56	4	93.33	6.67
LLaMa2	48	12	80.00	20.00
Mixtral	52	8	86.67	13.33
		Average	86	14

Through the evaluation of five LLMs (GPT-4, GPT-3.5, Llama-2, Claude2, and Mixtral-8x7B) on the clinical case reports from PMC-Patients dataset, the study reports that the accuracy of differential diagnoses improves substantially when lab test results are included, underscoring their critical role in accurate medical diagnosis. The inclusion of lab test results significantly enhances the accuracy and lenient accuracy of differential diagnosis predictions made by large language models, especially in improving the exact match predictions. Lab data, such as liver function tests, toxicology/metabolic panels, and serology/immune tests, were generally interpreted correctly, enhancing the models' ability to generate relevant diagnoses.

References

1. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 330, 78–80 (2023).
2. Berg, H. T. et al. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. Annals of Emergency Medicine 83, 83–86 (2024).

Acknowledgments

This work was supported by the AHRQ grant R21HS029969 (PI: ZH). QJ and ZL were supported by the NIH Intramural Research Program, National Library of Medicine.