AI-Driven System for Personalized Medical Diagnosis: Leveraging Llama for Dynamic Disease Prediction and Patient Data Integration

Francesco Boldrini
Alessandro Ciarniello

November 13, 2024

Contents

Chapter 1

Introduction

In recent years, research on artificial intelligence (AI) has focused on developing models capable of automating, either partially or fully, complex processes, making them more efficient in terms of both accuracy and time. AI is becoming an increasingly important tool in medicine, particularly for supporting clinical diagnosis to improve accuracy and speed. However, the complexity of the diagnostic processâĂŤwhich involves evaluating medical histories, symptoms, and test resultsâĂŤpresents significant challenges. Accurate and rapid diagnoses are crucial for reducing errors and improving patient care, especially in complex cases that require differential diagnosis (DDx). Currently, Large Language Models (LLMs) developed by institutions like Google Research and OpenAI (ChatGPT) are being used to support this process. These tools have shown promising results but remain limited in terms of the fluid, iterative interactions needed in real clinical settings.

Chapter 2

State of The Art

Recent studies show that both specialized medical LLMs and general-purpose models, like GPT-4, have achieved promising results in diagnostic accuracy and time to diagnosis. Two main articles [2][1] were referenced to understand and analyze the current state of the art, which detail results from experimental trials in medical diagnosis using LLMs.

2.1 Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial

This study was conducted to evaluate the effectiveness of a "General Purpose" language model (ChatGPT Plus, based on GPT-4) as a diagnostic tool in a clinical setting, comparing it to traditional research resources. The study included 50 doctors, divided into groups with and without LLM assistance, with performance assessed on six clinical vignettes.

2.1.1 Key Results

Diagnostic Average Score: This expresses an overall measure of the quality of a diagnosis made by a doctor or system. The average diagnostic score for doctors assisted by the LLM was 76% compared to 74% for those with traditional resources, with no significant improvement (a 2% difference; 95% Confidence Interval: -4% to 8%, p=0.60). This result indicates that the LLM did not significantly improve the overall performance of doctors compared to traditional resources.

Final Diagnosis Accuracy: Even when treating the result as a binary variable (correct vs. incorrect), the LLM showed slightly greater accuracy, but this was not statistically significant (OR 1.4, 95% Confidence Interval: 0.7 to 2.8, p=0.39).

Diagnosis Duration: The average time taken to complete each case was shorter for clinicians with LLM assistance (519 seconds) compared to those with traditional resources (565 seconds), but the difference was not statistically significant (p=0.20).

2.2 Towards Accurate Differential Diagnosis with Large Language Models

In this second experiment, a language model optimized for differential diagnosis (DDx) support was developed and evaluated on a series of complex clinical cases selected from the New England Journal of Medicine (NEJM). The study focused on comparing the performance of the LLM with that of doctors working without assistance or with the help of traditional research resources (such as Google or PubMed).

2.2.1 Key Results

Top-10 Accuracy The diagnostic accuracy, defined as the presence of the correct diagnosis among the top 10 suggestions, was significantly higher for the LLM (59.1%) compared to clinicians without assistance (33.6%) and those with traditional research resources (44.4%). This result suggests a clear improvement in diagnostic capability with the support of the LLM.

Quality of DDx: Doctors assisted by the LLM reported a higher quality of diagnosis, with an average score of **4.43** out of 5, compared to non-assisted clinicians **(3.74)** and those with research resource support **(3.80)**. Statistical testing showed significant differences between the LLM and the other conditions ($p<0.01$).

Comprehensive Score: The score for completeness of the DDx (including reasonable diagnostic candidates) was also higher for clinicians assisted by the LLM, with an average score of **4.06** compared to **3.80** for doctors using traditional research resources.

2.2.2 Conclusions

These results highlight how LLMs could represent a promising tool for improving the formulation of differential diagnoses. However, the optimal integration of these models into clinical workflows requires a design that leverages their capabilities without slowing down decision-making processes or generating ambiguity.

The results from the article *Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial*, for example, suggest that a commercial LLM chatbot (ChatGPT) might be more effective as an autonomous diagnostic tool. However, to

achieve effective results in comparison to a direct integration into the current diagnostic workflow of doctors, a dedicated system designed for the medical field would be necessary

Chapter 3

PROJECT DEVELOPMENT

The main objective of this study is to develop an application supported by a language model (LLM) to make the diagnostic process in the medical field more accurate and efficient. Specifically, the LLM is used as a tool for performing differential diagnosis of clinical cases, basing the analysis on both the patient's current health status and their medical history.

3.1 STATE-OF-ART ANALYSIS AND STARTING POINT

After analyzing the state of the art, paying particular attention to statistical results, it was evident that a successful application in this field requires higher diagnostic accuracy than that achieved in the trials of the two referenced studies. Therefore, the models used in these studies served both as a starting point for developing our project and as benchmarks for statistically analyzing our results.

3.2 LLM CHOICE AND DATA SELECTION

3.2.1 Llama

For the development of this application model, the LLM used is **Llama (Large Language Model Meta AI)**. Developed by Meta (formerly Facebook), Llama was designed as an advanced language model specifically trained for natural language processing (NLP). Currently, several versions of this Large Language Model have been released; for this project, the version used is *âĂIJ3.2-vision 90b q8âĂİ*. The possibility of using the latest versionâĂŹs image processing capabilities were also evaluated, as it

allowed the inclusion of images as input. However, this version is heavier to run with image classification and the resources available for this project did not allow for its full utilization.

Llama was run on a dedicated server equipped wit 4 4070 super graphics cards and 128 gigabytes of RAM, to ensure efficient and fast data processing. This configuration was chosen to leverage the computational power needed to handle large volumes of text. In fact, the server configuration provides the ability to run such an infrastructure to manage the computational load of the LLM.

3.2.2 DATASET

As previously mentioned, the results of the experiments reported in the articles cited in *Chapter 2* were used as a statistical benchmark. To ensure that the results would be meaningfully comparable, input data with the same structure as those used in the articles were sought.

Initially, the possibility of using the same dataset from one of the two experiments (which used different datasets but with a similar structure) was considered, but unfortunately, neither of them was available publicly.

Consequently, accessible datasets containing descriptions of clinical cases and their respective diagnoses were searched for.

The dataset selected for this project is: **MultiCaRe_Dataset**.

(Link to raw Dataset: https://github.com/mauro-nievoff/MultiCaRe_Dataset/tree/main)

The **MultiCaRe_Dataset** is a dataset of clinical cases, images, and captions from freely accessible case reports on PubMed Central. The dataset consists of about 34,000 cases, but not all of them contained the correct diagnosis. After filtering the cases with diagnoses (around 1,700), approximately 150 cases were selected based upon descripton length and relevance. This selection was made to analyze several clinical cases equal to those analyzed in the experiments, thus ensuring the statistical comparability of the results.

However, before providing the data to Llama for diagnosis, a series of pre-processing operations were performed, mainly focused on:

- **Text formatting correction:** modification of the separation characters between words and text lines; modification of the tabulation character
- **Diagnosis separation:** separating the actual diagnosis from the descriptive text of the clinical case, to avoid providing the correct answer to the LLM.

- **Stemming of diagnosis text:** used to avoid mismatch in the event of similar wording, both stemming techniques and custom dictionaries were used.

Chapter 4

SCRIPT FOR PROCESSING

Once the data were pre-processed, a Python script was created to quickly and fully automatically pass the descriptions of the various clinical cases to Llama, which processes them and generates the results.

* See github for the Scripts

Chapter 5

CONCLUSION

This study explored the application of Llama as a support for differential diagnosis in the medical field. It is expected that Llama could surpass other widely used models, such as ChatGPT and Med-PaLM, in accuracy, thereby providing more precise and reliable diagnoses. The comparative analysis with these tools will help identify the potential of Llama and the areas where it could be further optimized for specific clinical cases. **In particular, Llama can be run locally, avoiding the vast majority of problems linked with using remote tools in delicate fields such as the ones involving the management of sensitive medical informations.**

The use of Llama demonstrated the effectiveness of advanced technological configurations in handling large volumes of clinical data and supporting medical diagnoses with efficient processing times. Expectations for improved diagnostic accuracy with Llama suggest a significant step forward toward more effective integration of LLMs into clinical practice, with the goal of reducing diagnostic errors and supporting physicians in the decision-making process. Based on these results, future developments of the project may include further fine-tuning of Llama on specialized clinical datasets, improving its adaptability to complex and specific diagnostic contexts.

Our Results have shown an accuracy as follows:

Unsuggested (all zeros): 21.32% Well suggested (first match is 1): 55.88% Suggested (at least one match is 1): 78.68%

This has been counted by setting to 1 matches where diagnosis keywords were detected in each of the 5 possible cases, leaving to 0 any ill-formatted responses or otherwise misinterpreted answers (Thus suggesting our results are conservative and probably yield higher accuracy)

With an overall suggestion accuracy of almost 80%, our system even in its most primitive stage is promising and quite possibly efficient in aiding diagnostics in the field.

Chapter 6

MOCKUP IDEA

Finally, a potential interface was created for the application that would implement this model.

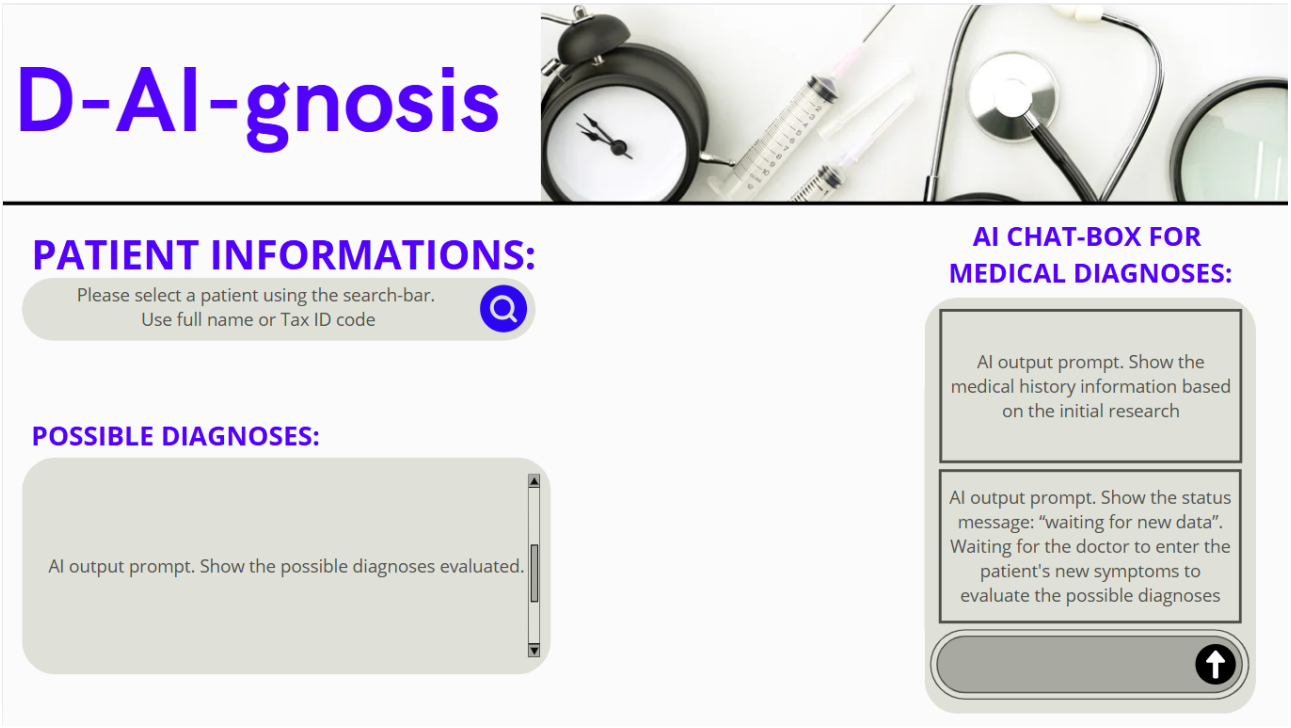

Figure 6.1: Mockup

The interface consists of 3 main elements:

Patient InformationâĂŹs search bar: The search bar allows the doctor to enter the personal information of the patient they are examining, to retrieve their medical history.

AI chat-box: Once the patient is found, their medical history (if available) is displayed

within the chat box. After displaying the medical history, a waiting message appears, indicating to the doctor that the LLM is awaiting the new symptoms of the patient to proceed with the diagnosis.

Possible diagnoses box: In this window, the possible diagnoses proposed by the LLM are displayed.

Bibliography

[1] Tao Tu Anil Palepu Amy Wang Jake Garrison Karan Singhal Daniel McDuff, Mike Schaekermann et al. Towards Accurate Differential Diagnosis with Large Language Models.

[2] Jason Hom Eric Strong Yingjie Weng Hannah Kerman JosÃľphine Cool Ethan Goh, Robert Gallo et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial.