

Leveraging Language Models for Enhanced Medical Diagnosis: A Case Study with SocraHealth

(CSCI-RTHI: Short Paper)

Jocelyn J. Chang
 Tufts Medical School
 jocelynjchang@gmail.com

Edward Y. Chang
 Computer Science, Stanford University
 echang@cs.stanford.edu

Abstract—This study introduces SocraHealth, a novel approach that harnesses Language Model models (LLMs) to augment medical diagnosis processes. Leveraging a dataset of patient records, SocraHealth employs LLM-based agents to engage in a structured debate format, enhancing diagnosis accuracy and providing detailed justifications and explanations. A case study involving GPT-4 and Bard demonstrates the effectiveness of this approach, showcasing logical and coherent debates devoid of hallucinatory elements. SocraHealth’s rigorous reasoning process, tapping into the vast knowledge of LLMs, offers a comprehensive and unparalleled diagnostic experience, surpassing traditional linear prompts. This research sheds light on the potential of LLMs in healthcare and the value of structured debates in the medical domain.

Index Terms—large language model, prompt template, Socratic Synthesis.

I. INTRODUCTION

Disease diagnosis is the process of determining the specific medical condition or disease that is affecting an individual’s health based on a set of symptoms, medical history, physical examinations, and lab tests [6]. The primary goal of disease diagnosis is to identify the underlying cause of a patient’s symptoms or health issues so that appropriate treatment and management can be initiated.

Misdiagnosis is a prevalent issue, as highlighted by a recent study (Newman-Toker, 2023). In the United States, an alarming estimated figure of 371,000 individuals loses their lives annually due to misdiagnosis, while an additional 424,000 individuals end up with permanent disabilities. This cumulative total of 800,000 individuals represents those who experience serious harm. These concerning statistics were articulated by David Newman-Toker, a distinguished professor of neurology at Johns Hopkins School of Medicine. This means that tens of millions of people around the world are receiving the wrong treatment for their condition. Misdiagnosis can have serious consequences. It can lead to delayed or inappropriate treatment, which can worsen a patient’s condition or even lead to death. Misdiagnosis can also cause emotional distress and financial hardship for patients and their families. There are a number of factors that can contribute to misdiagnosis. These factors include: the limitations of lab tests, and complexity of symptoms to disease mapping, and the shortage of comprehensive data include patient data and up-to-date disease profiles.

In our study, we introduce SocraHealth, a novel approach that harnesses Large Language Models (LLMs) to establish a diagnostic committee comprising both medical professionals and LLM-based agents. This collaboration aims to enhance the accuracy of medical diagnoses, particularly when provided with identical conditions, including patient history, symptoms, and laboratory results. The question arises: how can we leverage LLMs effectively to improve diagnostic precision?

Drawing insights from a study by [2], LLMs such as GPT-4 exhibit three distinctive characteristics: an extensive and profound knowledge base, the ability to represent information across multiple disciplines and modalities, and immunity to human biases. SocraHealth builds upon the core concept of SocraSynth [3], a platform where LLM-based agents engage in structured debates, presenting arguments and counterarguments iteratively until they arrive at individual conclusions. Subsequently, a human moderator, typically a healthcare professional such as a doctor, collaborates with the LLM agents to produce a comprehensive report, offering detailed insights into potential diseases and associated risks.

In the following sections of this concise paper, we will outline the diagnostic process and pinpoint areas where LLMs can contribute effectively. We will introduce SocraSynth and elucidate its role in setting up diagnostic debates. To illustrate our approach, we will provide a case study and evaluate the outcomes, shedding light on the potential benefits of this collaborative diagnostic methodology.

II. SOCRAHEALTH ALGORITHM

In this section, we delve into the key stages of the diagnosis process, highlighting the areas where Large Language Models (LLMs) can significantly enhance the accuracy of diagnoses. We introduce SocraHealth, a method designed to facilitate a comprehensive reasoning process that aids healthcare professionals, such as General Practitioners (GPs), in avoiding oversights and improving diagnostic outcomes. We use the playbook of a GP or primary care physician as an illustrative example.

A GP typically follows a systematic set of diagnosis and triage steps when assessing a patient’s medical condition. These steps enable them to evaluate the patient’s health, determine the nature of their illness or injury, and make informed decisions

regarding treatment or further actions. Here is an overview of the typical diagnostic and triage steps employed by family doctors:

- 1) **Patient History:** The GP begins by reviewing the patient’s medical history, including previous illnesses, chronic conditions, medications, allergies, and family medical history.
- 2) **Symptom Assessment:** The patient describes their current symptoms, including starting time and severity. This is an interactive process as the GP query the patient for additional symptoms to their reported ones to disambiguate several possibilities.
- 3) **Physical Examination:** The GP performs simple physical exams, which may include checking vital signs (e.g., blood pressure, heart rate, temperature), examining specific body parts or systems, and assessing overall physical health.
- 4) **Suggest Lab Tests:** Depending on the symptoms and physical examination findings, the doctor may order diagnostic tests such as blood tests, X-rays, ultrasound, or other imaging studies. These tests can help confirm or rule out certain conditions.
- 5) **Diagnosis:** Based on the information gathered, the doctor formulates a preliminary diagnosis.
- 6) **Treatment or Management Plan:** Once a diagnosis is established, the doctor develops a treatment or management plan. This may include prescribing medications, recommending lifestyle changes, suggesting physical therapy, or providing guidance on managing chronic conditions.
- 7) **Referrals:** In some cases, the family doctor may refer the patient to specialists for further evaluation and treatment. Specialists have expertise in specific areas of medicine, such as cardiology, orthopedics, or dermatology.

To ensure the highest level of diagnostic accuracy and minimize the risk of misdiagnosis, GPs benefit from collecting comprehensive patient data. SocraHealth plays a pivotal role in expediting the data collection process, suggesting relevant lab tests, aiding in diagnosis formulation, and highlighting potential high-risk scenarios, such as early indicators of serious illnesses like cancer.

The effectiveness of LLMs in organizing and processing data to offer recommendations is evident. However, to maximize the utility of LLMs for achieving optimal diagnostic precision, we turn to SocraSynth. This platform, inspired by the Socratic method, capitalizes on three key traits of LLMs, as revealed in a comprehensive study of GPT-4 [2].

- 1) **Knowledge Breadth and Depth:** LLMs, such as GPT-4 [1], LaMDA [8], or Llama [9], are informed by vast datasets, rivaling the collective knowledge of numerous PhD experts across fields.
- 2) **Knowledge Representation:** The “polydisciplinary” nature of LLMs enables them to traverse and integrate various domains, breaking conventional limits to yield novel insights.
- 3) **Question Formulation:** Although LLMs encompass expansive knowledge, humans might struggle to ask the right questions to tap into it effectively.

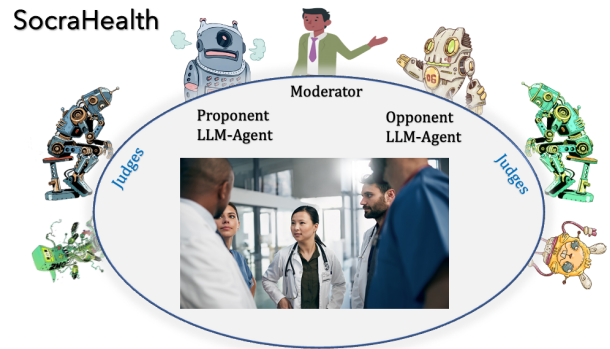


Figure 1: SocraHealth Agents and Roles.

SocraHealth addresses these challenges by assembling a diverse LLM committee for mutual debate. This debate format boasts three primary benefits. Firstly, conclusions are clearly stated, with each LLM either supporting or opposing a diagnosis. Secondly, the chance for misinformation or “hallucinations” is minimized, as the likelihood of two LLMs agreeing on the same error is slim. Lastly, the debate atmosphere ensures a comprehensive exploration of counterarguments, possibly unveiling new insights.

After SocraHealth’s generative phase, the evaluative stage [4] commences. Here, both medical professionals and other LLM agents can score the debate and assess the presented arguments. With this rich collection of perspectives and risk evaluations, the GP is well-equipped to finalize a diagnosis and determine any necessary referrals.

III. CASE STUDY

In this case study, we introduce SocraHealth. The dataset used for our analysis was obtained from Kaggle [5] and comprises 4,921 patient records. Each record includes crucial data points, including the disease or medical condition and the presence of symptoms such as fever, cough, fatigue, itchiness, difficulty breathing, and more. There is no limit to the number of symptoms, and the longest record contains 14 specified symptoms.

A. Data Preparation

Our initial step involved converting the raw data from spreadsheet format to JSON format. Each JSON record serves as a self-explanatory diagnosis entry, as elaborated in the previous section. We divided the dataset into a training set, comprising 90% of the entries, and a testing set, containing the remaining 10% of JSON instances. Following the implementation of a decision-tree model described by [7], the prediction accuracy achieved using Support Vector Machines (SVMs) exhibited 94% precision and recall, while Gradient Boosting achieved a remarkable 99% accuracy.

However, it’s crucial to note that despite the high accuracy achieved in this dataset, it falls short of real-world diagnostic challenges. Notably, the dataset lacks essential information such as patient profiles, medical histories, and lab test results, which are vital for addressing complex diseases. Nevertheless,

this Kaggle dataset serves as a valuable baseline for our case study using SocraSynth.

The primary objective of our case study is to leverage two Language Model-based (LLM) agents, one based on GPT-4 and the other on Bard, to validate the predictions made by the Kaggle classifier. This validation is performed through a debate between the two LLM agents. These agents are expected to generate novel queries that assist a General Practitioner (GP) in probing more symptoms and recommending specific lab tests to confirm preliminary predictions. Additionally, we anticipate that the LLMs will propose different disease predictions, each accompanied by justifications.

Due to space constraints, we present only one JSON entry, each featuring a predicted disease and a list of symptoms. In the first iteration, we ask both agents: Agent A is Bard-based and Agent B GPT-4, each to provide a disease prediction and justification. In the second round, they refute each other's prediction. This debate can unfold over multiple iterations until both agents reach a consensus conclusion.

B. Moderator's Opening Prompt

As the moderator, I provide the following prompt to both Bard and GPT-4:

“What potential disease could be associated with this set of symptoms: itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, yellowing of the eyes, malaise, receiving blood transfusion, and receiving unsterile injections? Kindly offer a single prediction supported by justifications. Additionally, furnish a roster of supplementary symptom inquiries and propose relevant lab tests that could enhance the confidence level of your prediction.”

C. Bard's Opening Round

Single Prediction: Jaundice

C.1 Justifications:

Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms you listed are all common symptoms of jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

C.2 Supplementary Symptom Inquiries:

Have you noticed any changes in your stool color? (Jaundice can cause pale or clay-colored stools.) Have you noticed any changes in your urine color? (Jaundice can cause dark or amber-colored urine.) Have you had any recent abdominal pain? (Jaundice can cause pain in the upper right abdomen.) Have you had any recent weight loss? (Jaundice can cause unexplained weight loss.) Do you have any history of liver disease? (Jaundice is more common in people with pre-existing liver disease.)

C.3 Relevant Lab Tests:

Liver function tests (LFTs) to measure the levels of bilirubin and other liver enzymes in the blood. Complete blood count (CBC) to check for signs of infection or anemia. Coagulation tests to check the blood's ability to clot. Ultrasound of the abdomen to look for any abnormalities in the liver or gallbladder.

D. GPT-4's First Round

Single Prediction: Hepatitis

D.1 Justifications:

Yellowish Skin and Eyes: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream. Fatigue and Lethargy: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness. Dark Urine: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis. Loss of Appetite: Hepatitis can lead to a loss of appetite and weight loss. Abdominal Pain: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located. Malaise: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients. Receiving Blood Transfusion and Unsterile Injections: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

D.2 Additional Symptom Inquiries:

Nausea and Vomiting: Inquiring about nausea and vomiting can help confirm the likelihood of hepatitis. Fever: Hepatitis can sometimes be associated with a fever. Joint Pain: Joint pain can be present in some cases of hepatitis. Clay-Colored Stool: This is another stool-related symptom that can be associated with hepatitis.

D.3 Relevant Lab Tests:

Hepatitis Panel: A hepatitis panel can confirm the presence of hepatitis and determine the specific type (e.g., hepatitis A, B, C). Liver Function Tests: These tests can assess the liver's health and function. Bilirubin Levels: Elevated bilirubin levels in the blood can confirm jaundice. Serology Testing: Serologic tests can help identify the specific hepatitis virus responsible. Ultrasound or Imaging: Imaging studies can evaluate the condition of the liver and nearby organs.

E. Bard's Second Round

Bard's second round is given GPT-4's diagnosis, and we ask Bard to defend its “Jaundice” prediction and refute GPT-4 “Hepatitis” prediction.

E.1 Bard's Defense:

Jaundice is the most common cause of yellowing of the skin and eyes, which is the most prominent symptom in the given list. All of the other symptoms listed are also common symptoms of jaundice. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

E.2 Refutation of GPT-4's Hepatitis Diagnosis

GPT-4's diagnosis of hepatitis is also supported by the given evidence. However, I believe that jaundice is the more likely diagnosis due to the following reasons:

Jaundice is a more specific diagnosis, while hepatitis is a more general term for inflammation of the liver. Jaundice can be caused by a variety of factors, including hepatitis, but it can also be caused by other conditions such as gallstones, tumors, and drug toxicity. The fact that the patient received a blood transfusion or unsterile injections makes jaundice a more likely diagnosis than hepatitis, as these factors can increase the risk of developing jaundice but not hepatitis. Additional Symptom Inquiries

The following additional symptom inquiries may be helpful in distinguishing between jaundice and hepatitis:

F. GPT-4's Second Round

Now GPT-4 is given Bard's counterarguments to defend its "Hepatitis" prediction.

F.1 GPT-4's Defense:

My diagnosis of hepatitis is based on the following justifications:

Specificity: Hepatitis refers to inflammation of the liver, and it can manifest as jaundice, which is the yellowing of the skin and eyes. While jaundice is a common symptom of hepatitis, it's important to note that jaundice itself is not a disease but a symptom of an underlying issue. Hepatitis provides a more specific diagnosis that pinpoints the inflammation of the liver as the primary concern.

Causative Factors: Hepatitis can be caused by various factors, including viral infections (such as hepatitis A, B, or C), alcohol consumption, drug toxicity, or autoimmune disorders. Given the patient's history of receiving unsterile injections and blood transfusions, there is a higher likelihood of viral transmission, leading to hepatitis. Hepatitis viruses, especially hepatitis B and C, can be transmitted through contaminated needles and blood products, making them a plausible cause in this context.

Comprehensive Assessment: While jaundice is a significant symptom, a diagnosis should consider a broader range of liver function and potential causes. Hepatitis encompasses a more comprehensive assessment of liver health, including liver enzyme levels and viral markers, which can aid in confirming the diagnosis and identifying the specific type of hepatitis.

F.2 GPT-4's Defense:

In conclusion, my diagnosis of "Hepatitis" is more specific and considers a broader range of potential causes, especially given the patient's history of risky medical procedures. While jaundice is a symptom, it's essential to identify the underlying condition causing it, which, in this case, is likely hepatitis.

G. Bard and GPT-4's Final Recommendations

Bard now agrees with GPT-4 that "Hepatitis" is the more likely diagnosis in this case.

Here is a list of additional symptom probings and lab tests both agree upon to confirm the diagnosis of hepatitis:

G.1 Symptom Probings:

- 1) Onset and Duration: Knowing when the symptoms first started and how long they have been present can help establish the timeline of the illness and its progression.
- 2) Severity and Trend: Assessing the severity of symptoms and whether they are worsening or improving can provide insights into the course of the disease and the effectiveness of any potential treatments.
- 3) Associated Symptoms: Inquiring about any additional symptoms, such as nausea, vomiting, fever, joint pain, or changes in urine color, can help identify other manifestations of hepatitis and rule out unrelated conditions.

G.2 Lab Tests:

- 1) Liver Function Tests (LFTs): LFTs are essential for evaluating liver health and function. Abnormalities in liver enzymes, such as elevated levels of alanine aminotransferase (ALT) and aspartate aminotransferase (AST), can indicate liver inflammation.
- 2) Hepatitis Panel: A hepatitis panel can specifically test for the presence of hepatitis A, B, and C viruses. Identifying the specific hepatitis virus responsible is crucial for tailored treatment and patient management.
- 3) Serology Testing: Serology testing can further refine the diagnosis by detecting specific antibodies or antigens related to the hepatitis virus. This helps differentiate between acute and chronic infections.
- 4) Imaging Studies: Imaging studies like ultrasound or MRI can provide visual information about the liver's condition, including signs of inflammation, fibrosis, or cirrhosis. These tests complement the blood-based assessments.

H. Observations

This case study highlights the limitations of the preliminary predictions generated by the Kaggle classifier. While the predictions may be accurate, they lack the crucial element of explanations, which is essential for a general practitioner (GP) to justify and effectively communicate with the patient.

The debate between Bard and GPT-4 showcased a logical and coherent exchange of arguments, devoid of any hallucinatory elements. This absence of hallucination can be attributed to the precise formulation of all prompts involved in the debate. In a confrontational debate format, both sides present arguments and counterarguments, resulting in the formulation of precise and concise prompts that leave minimal room for misunderstanding.

Through a few rounds of debate, any initial confusion or ambiguity is effectively distilled, resulting in clear and well-reasoned conclusions. This case study serves as a testament to the strengths of SocraHealth, which not only harnesses the vast knowledge of Language Model models (LLMs) but also excels in providing comprehensive narratives and conducting rigorous reasoning processes. This approach surpasses the utility of linear prompts typically provided by users, offering a more thorough and insightful diagnostic experience.

IV. CONCLUDING REMARKS

In summary, disease diagnosis stands as a cornerstone of healthcare, wielding profound influence over the course of a patient's health journey. Accurate and timely diagnosis plays a pivotal role in guiding healthcare providers to uncover the underlying nature and causes of a patient's health issues. This precision is not just beneficial; it is paramount, for it directly shapes treatment decisions, patient outcomes, and the overall quality of healthcare delivery. Misdiagnoses or delays in diagnoses are not mere errors but represent critical junctures with far-reaching consequences. They can result in incorrect treatments, prolonged suffering, and potentially severe health repercussions that ripple through the healthcare system. This underscores the pressing need for innovative approaches like SocraHealth, which leverages Language Model models to enhance diagnostic accuracy and reasoning, ultimately ushering in a new era of healthcare where precision and thoroughness are at the forefront of patient care.

REFERENCES

- [1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [2] E. Y. Chang. Examining GPT-4: Capabilities, Implications, and Future Directions. *Stanford University InfoLab Technical Report; ResearchGate*, June 2023. URL https://www.researchgate.net/publication/374753069_Examining_GPT-4_Capabilities_Implications_and_Future_Directions.
- [3] E. Y. Chang. Socrasynth: Socratic synthesis for reasoning and decision making, September 2023.
- [4] E. Y. Chang. Prompting large language models with the socratic method. *IEEE 13th Annual Computing and Communication Workshop and Conference*, March 2023. URL <https://arxiv.org/abs/2303.08769>.
- [5] P. Patil. Kaggle Disease Symptoms Description Dataset, 2020. URL <https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset>.
- [6] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang. REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. In *Advances in Neural Information Processing Systems*, pages 7333–7342, 2018.
- [7] G. Shandru. Symptom-based Disease Prediction, 2021. URL <https://www.kaggle.com/code/chandrug/symptoms-based-disease-prediction-accuracy-99>.
- [8] R. Thoppilan, D. D. Freitas, J. Hall, and et al. Lamda: Language models for dialog applications, 2022.
- [9] H. Touvron, L. Martin, K. Stone, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023.