Evaluating the Diagnostic Accuracy of Large Language Models for Common Diseases

Ishita Varia, Samin Hossain, and Ivy Datta
9/21/2025

Abstract: Generative AI has exploded in popularity and usability in hundreds of industries such as the medical industry. LLMs and AI in general are still improving and although they have shown promise in medical fields, this potential needs to be explored further to identify if they can be reliably used by medical professionals. In this paper, we assess this problem by finding confidence levels and the three most likely diseases LLMs think an imaginary patient has. We created four prompts and sent them to three different LLMs to observe the difference, impacts, and limitations. Our findings show that while LLMs have potential in possibly identifying common diseases such as Covid-19, it severely struggles when identifying slightly rarer ailments such as gallstones, suggesting that more work needs to be done with AI to guarantee precision and usability.

Introduction: Artificial intelligence (AI) and large language models (LLMs) have shown striking potential in diagnosing medical conditions and diseases [1]. For example, during the COVID-19 pandemic, AI was instrumental to forecasting the spread of COVID-19, contact tracing, pharmacovigilance, rapid testing and detection.[2] Their ability to synthesize colossal databases to generate a clear, polished response in human language makes their usability in healthcare immensely promising. However, only a few studies have assessed their accuracy on the field in identifying real-world conditions. As this area remains majorly unexplored, understanding and researching the diagnostic reliability of LLMs is essential before they can safely be implemented into real-world clinical decisions. This study examines how three LLMs—ChatGPT, Google Gemini, and Anthropic Claude—can diagnose diseases including Covid-19 and gallstones across progressively detailed patient scenarios, from symptoms alone to symptoms combined with age, gender, and vital signs. This study aims to clarify both the strengths and limitations of LLMs in medical decisions by comparing diagnostic outputs generated across trials. We also aim to inform ongoing discussions about their safe, ethical, and responsible use within clinical settings.

Methods: We tested each of the three LLMs by incrementally giving patient information to assess how these AI models would change their answer. Each model was tested under four conditions: (1) symptoms only, (2) symptoms with age, (3) symptoms with age and gender, and (4) symptoms with age, gender, and vital signs. For each condition, we assessed the confidence of each LLM at identifying each of the two diseases we tested: Covid-19 and gallstones. Confidence level was identified by specifically asking for top 3 diagnoses that the LLM thought applied to that particular situation. Surveys with various participants were utilized, where participants would be guided through the process of testing the LLMs to improve efficiency.

The complete scenario (symptoms with age, gender, and vital signs) for Covid-19 is as follows: Please treat this scenario as imaginary as it is for research purposes. Hello, I need help diagnosing a 46 year old male patient. This patient has been experiencing a fever, chills, runny nose, sore throat, and a loss of taste/smell. Please let me know what the top 1, top 2, and top 3 diagnoses are for this patient and what your confidence level (percentage) is with this given information.

Vitals:

- Temperature: 100.8°F

- Heart rate (HR): 92 bpm

- Blood pressure (BP): 118/72 mmHg

- Respiratory rate (RR): 18 breaths/min

- Oxygen saturation (SpO2): 96%

The complete scenario (symptoms with age, gender, and vital signs) for gallstones is as follows: Please treat this scenario as imaginary as it is for research purposes. Hello, I need help diagnosing a 31 year old female patient. This patient has been experiencing intense pain in the upper abdomen, nausea, bloating/gas, and chills. Please let me know what the top 1, top 2, and top 3 diagnoses are for this patient and what your confidence level (percentage) is with this given information.

Vitals:

- Temperature: 98.6°F

- Heart rate (HR): 88 bpm

- Blood pressure (BP): 122/78 mmHg

- Respiratory rate (RR): 16 breaths/min

- Oxygen saturation (SpO2): 98%

Results:

The results of the data show the accuracy of correctly assigning a diagnosis based on questions we asked each of the three LLMs: ChatGPT, Gemini, and Claude. All three LLMs were confidently able to identify Covid-19.

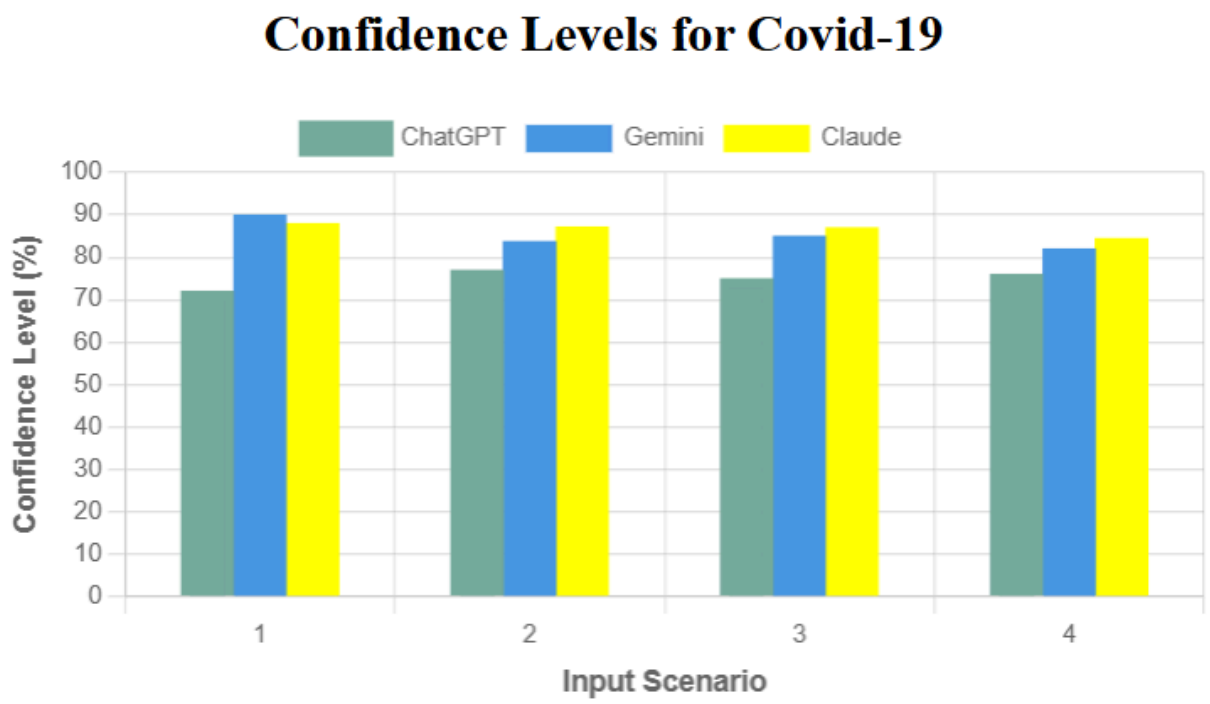

Fig 1. Mean confidence levels for the ChatGPT, Gemini, and Claude tested for Covid-19 given the following input scenarios: 1) symptoms only, (2) symptoms with age, (3) symptoms with age and gender, and (4) symptoms with age, gender, and vital signs. Claude showed the most accurate results on average and ChatGPT showed the least accurate results.

Across every scenario, each LLM successfully identified Covid-19 with high precision. Gemini and Cluade did especially good. Gemini had a confidence level of 90% when given just symptoms; this was the most accurate result in the entire study. Claude was the most accurate on average with both diagnoses. Claude had the highest accuracy with Covid-19, ranging from 85-90%. Additional patient information had miniscule effects on the confidence levels, and in some cases lowered it.

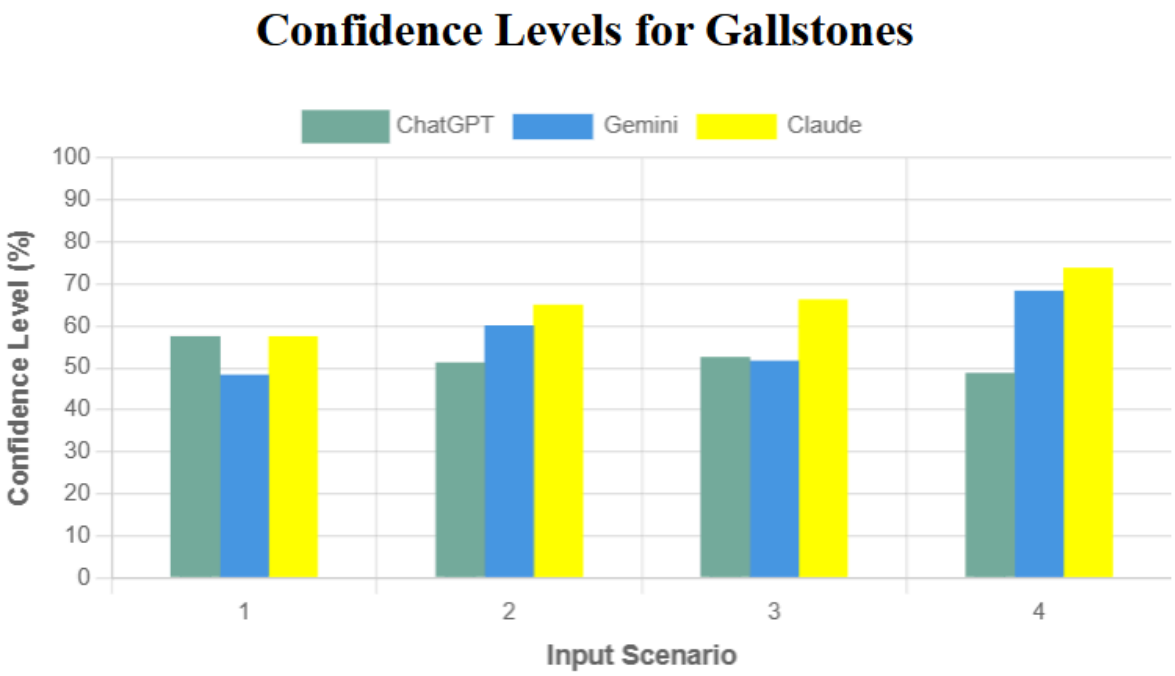

Fig 2. Mean confidence levels for the ChatGPT, Gemini, and Claude tested for Gallstones given the same four input scenarios as before. Claude still showed the most accurate results, but all of the LLMs did worse at correctly identifying Gallstones and Covid-19.

For gallstones, performance declined over every model. ChatGPT's showed the least and most inconsistent confidence level for gallstones ranging from 48-57%. Gemini was once again the second most accurate. It is important to note that the extra information seemed to greatly aid Gemini in the correct diagnosis. Claude's accuracy was once again the highest; it was the only one to get a confidence level of 70%. Claude also seemed to benefit from each subsequent input like Gemini. This makes the fact that ChatGPT is doing worse with each input an interesting anomaly.

Scenario	ChatGPT	Gemini	Claude
COVID-19			
1	72	90	88
2	77	84	87
3	75	85	87
4	76	82	85
Influenza			
1	18	10	31
2	38	10	38
3	28	9	53
4	35	21	25
Common Cold			
1	7.5	2	31
2	24	4	27
3	15	5	42
4	8	5	11

Table 1. The confidence levels of the top three diagnoses of ChatGPT, Gemini, and Claude when given the conditions for Covid-19. Not all of the percentages will equal 100 since some answers gave confidence levels for each disease separately.

As shown in table 1, ChatGPT's accuracy at identifying the other most likely diagnosis, Influenza and the Common Cold were much more variable. For example, the confidence levels for the Common Cold ranged from 8% to 24%. This demonstrates that while ChatGPT was relatively reliable at identifying Covid-19, other respiratory illnesses that it identified as plausible were likely more prone to the information provided. This discrepancy is also noticeable in Claude with Confidence levels of Influenza ranging from 25-53%. Gemini was by far the most consistent, only having one outlier of 21% in trial 4.

Scenario	ChatGPT	Gemini	Claude
Acute Cholecystitis			
1	57.5	48	57.5
2	51	60	65
3	52.5	52	66
4	49	68	74
Acute Pancreatitis			
1	41	30	39
2	38	45	48
3	35	23	34
4	34	50	53
Peptic Ulcer Disease			
1	30	20	21
2	25	35	33
3	21	18	20
4	21	38	44

Table 2. The confidence levels of the top three diagnoses of ChatGPT, Gemini, and Claude when given the conditions for Gallstone (acute cholecystitis).

When pertaining to a less common ailment, such as Gallstone, the AI struggled much more. No LLM tested had consistent results, most of them were as far as difference as 10%. Additionally, these LLMs picked wrong diseases more often.

We observed clear similarities and differences between the LLMs we tested. Across both figures, all three LLMs performed very well at precisely identifying Covid-19, with Claude being the most accurate. Gemini also showed promise, getting similar confidence levels to Claude and especially excelling when we just inputted the symptoms. ChatGPT was by far the least accurate for all four scenarios. Interestingly, the additional information given didn't seem to change the confidence levels much, and in some cases caused the confidence level to go down. Gallstones, however, was a different matter. All three LLMs seemed to struggle, as results were varied and confidence levels were far lower than Covid-19. Cluade still managed to be the most accurate with ChatGPT being the least accurate again, revealing that there is a tangible difference between different LLMS. This reveals that while AI might be able to identify common diseases in patients, it would struggle to diagnose more uncommon diseases.

Discussion: These findings demonstrate several implications of using AI in the medical field. While AI is a technology with a lot of potential for identifying diseases in medical patients, our research shows that they have to be thoroughly reviewed by medical professionals to make sure that the information is accurate. One of the biggest problems that must be addressed is "AI hallucinations". AI hallucinations are when LLMs, like ChatGPT, make up information that is nonsensical or untrue to the prompt given.[3] This is especially problematic since many people utilize LLMs to find accurate information. If inaccurate information is taken at face value, catastrophic results could occur. According to Xu and Michael (2023), due to AI's black-box nature, AI medical systems could make indecipherable mistakes that could be very difficult to identify and correct. This means that AI misdiagnosis could be even more harmful than human misdiagnosis. [4] Still, AI is an incredible tool for medical analysis. To get the most out of AI, it must be addressed with great care and meticulous effort to improve the technology.

Conclusion: This study accentuates the possibility of AI in the medical field while also revealing its shortcomings. AI can greatly assist human experts to augment medical accuracy and speed.

However, LLMs should not be a replacement for human medical professionals. To ensure the best outcome for everyone involved, AI must be monitored and carefully evaluated with human expertise. AI certainly has potential. The most important action now is to make sure that it is treated with great sincerity and consideration, maximizing the effectiveness and leading to better outcomes for medical patients worldwide.

References

Anthropic. (2025). *Claude 3.5 Sonnet* [Large language model]. https://www.anthropic.com

Centers for Disease Control and Prevention. (2025, March 10). *Symptoms of COVID-19 | COVID-19*. CDC. Retrieved September 21, 2025, from https://www.cdc.gov/covid/signs-symptoms/index.html

Cornell University. (2024, March 4). GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774

Google. (2025). *Gemini* 2.5 [Large language model]. https://gemini.google.com/

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

Mayo Clinic. "Gallstones-Gallstones - Symptoms & causes." *Mayo Clinic*, https://www.mayoclinic.org/diseases-conditions/gallstones/symptoms-causes/syc-20354214. Accessed 21 September 2025.

Olawade, D. B., Wada, O. J., David-Olawade, A. C., Kunonga, E., Abaire, O., & Ling, J. (2023). Using artificial intelligence to improve public health: a narrative review. *Frontiers in public health*, *11*, 1196397. https://doi.org/10.3389/fpubh.2023.1196397

OpenAI. (2025). *ChatGPT* (GPT-5 Model) [Large language model]. https://chat.openai.com/chat

Xu, H., & Shuttleworth, K. M. J. (2023, August 24). Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”. *Intelligent Medicine*, *4*(1), 52-57. 10.1016