Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241265215
journals.sagepub.com/home/dhj
Sage

Diagnostic performance of generative artificial intelligences for a series of complex case reports

Takanobu Hirosawa[1], Yukinori Harada[1], Kazuya Mizuta[1], Tetsu Sakamoto[1], Kazuki Tokumasu[2] and Taro Shimizu[1]

Abstract

Background: Diagnostic performance of generative artificial intelligences (AIs) using large language models (LLMs) across comprehensive medical specialties is still unknown.

Objective: We aimed to evaluate the diagnostic performance of generative AIs using LLMs in complex case series across comprehensive medical fields.

Methods: We analyzed published case reports from the *American Journal of Case Reports* from January 2022 to March 2023. We excluded pediatric cases and those primarily focused on management. We utilized three generative AIs to generate the top 10 differential-diagnosis (DDx) lists from case descriptions: the fourth-generation chat generative pre-trained transformer (ChatGPT-4), Google Gemini (previously Bard), and LLM Meta AI 2 (LLaMA2) chatbot. Two independent physicians assessed the inclusion of the final diagnosis in the lists generated by the AIs.

Results: Out of 557 consecutive case reports, 392 were included. The inclusion rates of the final diagnosis within top 10 DDx lists were 86.7% (340/392) for ChatGPT-4, 68.6% (269/392) for Google Gemini, and 54.6% (214/392) for LLaMA2 chatbot. The top diagnoses matched the final diagnoses in 54.6% (214/392) for ChatGPT-4, 31.4% (123/392) for Google Gemini, and 23.0% (90/392) for LLaMA2 chatbot. ChatGPT-4 showed higher diagnostic accuracy than Google Gemini ($P < 0.001$) and LLaMA2 chatbot ($P < 0.001$). Additionally, Google Gemini outperformed LLaMA2 chatbot within the top 10 DDx lists ($P < 0.001$) and as the top diagnosis ($P = 0.010$).

Conclusions: This study demonstrated the diagnostic performance of generative AIs including ChatGPT-4, Google Gemini, and LLaMA2 chatbot. ChatGPT-4 exhibited higher diagnostic accuracy than the other platforms. These findings suggest the importance of understanding the differences in diagnostic performance among generative AIs, especially in complex case series across comprehensive medical fields, like general medicine.

Keywords

Natural language processing, large language models, diagnostic accuracy, decision support system, diagnostic errors

Submission date: 22 March 2024; Acceptance date: 13 June 2024

Introduction

Diagnosis is fundamental to the practice of medicine. Pursuing diagnostic excellence is crucial for enhancing the quality and outcomes of medical care. Diagnostic excellence encompasses a holistic approach that ensures diagnoses are not only accurate and precise, but are also delivered in a manner that is safe, effective, patient-centered, timely,

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, Japan
[2]Department of General Medicine, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Corresponding author:
Takanobu Hirosawa, Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, 880 Kitakobayashi, Mibu-cho, Shimotsuga, Tochigi 321-0293, Japan.
Email: hirosawa@dokkyomed.ac.jp

efficient, and equitable.[1] This concept underscores the imperative of minimizing errors,[2] a critical step in optimizing the use of resources. Furthermore, it emphasizes the necessity to respect patients' unique needs and experiences throughout the diagnostic journey. Achieving diagnostic excellence necessitates the implementation of various strategies. Promoting a team-based diagnostic process emerges as particularly critical among these strategies. This strategy stands out due to its collaborative nature, thereby improving diagnostic accuracy and comprehensiveness through collective reasoning.[3] This collaborative approach involves not only healthcare providers, patients, and their families, but also integrates an essential tool: Clinical Decision Support Systems (CDSSs).[4]

The primary design goal of CDSSs is to enhance healthcare quality by refining the medical decision-making process. They play an impressive role in enhancing overall patient care.[5] CDSSs offer a multitude of capabilities, including diagnostic support. Representative examples of CDSSs underscore the diverse range of applications available. Examples of these tools include symptom checkers[6] and differential-diagnosis (DDx) generators,[7] such as DXplain.[8] Symptom checkers are online tools that yield a list of possible conditions from reported symptoms. DDx generators provide physicians with potential diagnoses based on collected medical information. DDx in the context of clinical practice denotes the comprehensive list of potential diagnoses compiled to account for a patient's symptoms,[9] a process streamlined by advanced DDx generators. DXplain, one of the DDx generators, employs artificial intelligence (AI) to analyze patient data and provide possible diagnoses based on clinical features. Despite their potential, several factors have limited the impact of CDSSs on clinical diagnostics. These factors include persistent negative perceptions among physicians and issues with accuracy. Another hurdle is the lack of seamless system integration,[5] which is essential for the effortless incorporation of CDSSs into existing healthcare workflows and electronic health record systems.

The limitations inherent in current CDSSs have necessitated the exploration of innovative solutions. Emerging as a promising response to these challenges is generative AIs, especially large language models (LLMs).[10] Generative AIs' ability to process extensive amounts of data with human feedback, understand nuanced human language, and generate precise, contextually relevant responses overcome some of the critical drawbacks of earlier systems. These AIs hold the potential to enhance diagnostic accuracy, personalized treatment recommendations, and improve patient engagement, addressing the gaps in existing CDSSs' accuracy and adaptability. In this innovative domain, several LLM-based chatbots have gained prominence as notable generative AI tools,[11] including the chat generative pre-trained transformer (ChatGPT) developed by OpenAI,[12] Google Gemini,[13,14] and LLM meta AI 2 (LLaMA2), a product of meta AI.[15]

These new generative AI technologies are poised to address pivotal questions in DDx.[16] Several studies have extensively researched the clinical potential of generative AI tools like the fourth-generation ChatGPT (ChatGPT-4).[17–19] Our investigation, among others, has shown that ChatGPT-4 exhibits more impressive diagnostic accuracy than its predecessor, the third-generation ChatGPT, ChatGPT-3.[20,21] Other studies indicated ChatGPT-4's superiority over other generative AIs in specialized fields, such as neurosurgery question banks,[20] respiratory medicine questions,[22] and general internal medicine (GIM) case series.[23]

Despite the promise shown by ChatGPT-4 and other generative AIs, few studies have comparatively analyzed their diagnostic performance. This gap exists because of certain tools, including Google Gemini and LLaMA2 chatbot utilizing the LLaMA2 model, remain under-evaluated. Additionally, their diagnostic performance across comprehensive medical specialties is still unknown. Although some studies have demonstrated their efficacy in comprehensive medical exams,[19,24] evaluations based on case reports, which are closer to clinical applications, have yet to be fully understood. This can be attributed to the intrinsic complexity and specialized nature of different medical fields,[25] which pose challenges for conducting broad-scale evaluations.

Therefore, this study aims to fill this gap in the current literature by comparing the diagnostic performance of ChatGPT-4, Google Gemini, and LLaMA2 chatbot for case series across a comprehensive range of medical specialties. By offering a comprehensive assessment across these diverse areas of medicine, this study not only seeks to fill a current gap in the literature but also to guide future development and applications of generative AIs toward enhancing diagnostic excellence.

Materials and methods

We conducted an experimental study to assess diagnostic accuracy of generative AI systems using data from published case reports. This study was conducted by the Department of GIM (Diagnostic and Generalist Medicine) at Dokkyo Medical University in Tochigi, Japan. As the research utilized published case reports, ethics committee approval was not applicable. The study involved preparing case materials, generating DDx lists utilizing AIs, and evaluating these lists. Figure 1 illustrates the study flow.

Preparing case materials

To provide a detailed understanding of the selection criteria for case reports, we included specific inclusion and exclusion criteria to ensure the generalizability of our findings. The case reports were sourced from the *American Journal*

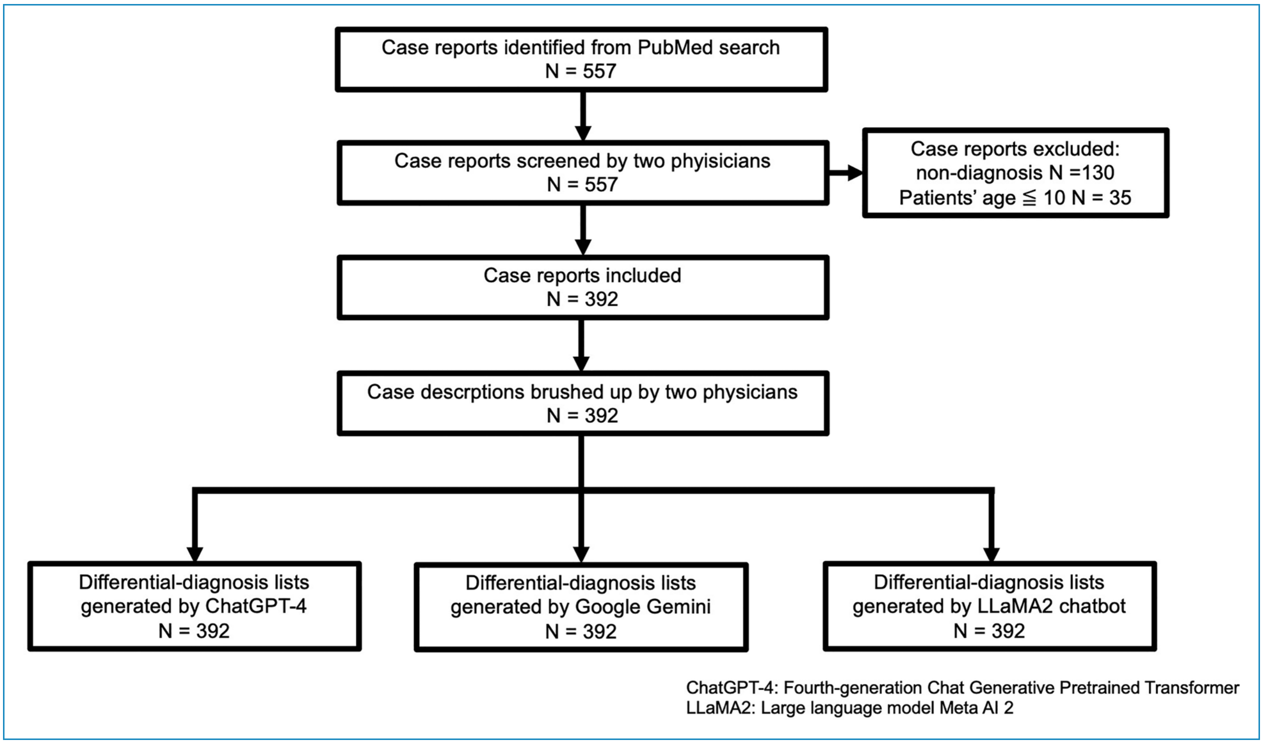

Case reports identified from PubMed search
N = 557
Case reports screened by two physicians
N = 557
Case reports excluded:
non-diagnosis N =130
Patients' age ≦ 10 N = 35
Case reports included
N = 392
Case descrptions brushed up by two physicians
N = 392
Differential-diagnosis lists
generated by ChatGPT-4
N = 392
Differential-diagnosis lists
generated by Google Gemini
N = 392
Differential-diagnosis lists
generated by LLaMA2 chatbot
N = 392
ChatGPT-4: Fourth-generation Chat Generative Pretrained Transformer
LLaMA2: Large language model Meta AI 2

Figure 1. Flowchart in the study.

of Case Reports. This peer-reviewed scientific journal publishes original, often complex, case reports spanning a variety of medical fields. Due to their structured nature, extracting case descriptions from them is straightforward. All case reports included in this study contained the patients' final diagnoses. We conducted a PubMed search for our sample selection, using the following keywords: "(2022/1/1:2023/3/1[dp])" AND "(American Journal of Case Reports[journal])." In the PubMed search, [dp] indicates the publication date, and [journal] denotes the specific journal. The inclusion criterion was case reports published by the *American Journal of Case Reports* between 1 January 2022 and 1 March 2023. The exclusion criteria were case reports primarily focused on management issues without specific diagnostic details and case reports involving patients younger than 10 years old. These exclusion criteria were based on a previous study on case reports that evaluated CDSS.[26] We prioritized the most recently published case reports. This was due to the 2021 knowledge cutoffs for generative pre-trained transformer 4, the model utilized in ChatGPT-4.[12] We initially identified 557 consecutive case reports, all of which were subjected to an eligibility review conducted by the main investigator (TH). Another investigator (YH) verified these processes. TH and YH excluded 130 articles because of their primary focus on management issues, and 35 were excluded due to involving patients younger than 10. For case reports documenting multiple cases, we included only the first case presented. After excluding 165 case reports, we included 392 case reports.

We edited these reports to highlight the case descriptions and final diagnoses. Typically, the case descriptions were extracted from the case report section of each article, which often detailed patient history, physical examinations, investigation results, and management prior to the final diagnosis. Other sections like titles, introductions, discussions, conclusions, and assessments present in the case report section, and any accompanying tables or figures were omitted. Any disagreements between TH and YH were discussed until a resolution was reached.

As an example, consider the case reports titled "A 75-Year-Old Woman with a 5-Year History of Controlled Type 2 Diabetes Mellitus Presenting with Polydipsia and Polyuria and a Diagnosis of Central Diabetes Insipidus,"[27] which was referenced as case number 1 in Supplementary Tables 1–3. From this article, we extracted case descriptions: from "A 75-year-old Japanese woman visited a primary care doctor due to a 2-month history of thirst, polydipsia…" to "The plasma AVP level was relatively low (0.6 pg/mL)." The final diagnosis for this case was central diabetes insipidus.

Generating differential-diagnosis lists utilizing artificial intelligences

In this study, we employed three different generative AIs: ChatGPT-4, Google Gemini, and LLaMA2 chatbot. None of the generative AIs, including ChatGPT-4, Google Gemini, and LLaMA2 chatbot, received additional training or reinforcement for medical diagnoses. Table 1 provides the details of the generative AIs employed in this study.

We utilized ChatGPT-4, a generative AI developed by OpenAI. This model, known for its advanced generation and comprehensive capabilities, is the GPT-4 series. We employed the 24th May version of ChatGPT-4 to generate DDx lists. In the version, there was no Internet connection option. ChatGPT-3.5 was excluded from this study due to its comparatively lower medical performance as observed in prior research.[28,29] Specifically, ChatGPT-4 demonstrated a significantly higher accuracy, surpassing ChatGPT-3.5. Given our aim to evaluate the most advanced generative AI systems, we selected ChatGPT-4. Figure 2A presents an example output of DDx list generated by ChatGPT-4 for a sample case description. Figure 2B explains Figure 2A. Supplementary Table 1 lists the DDx generated by ChatGPT-4 alongside the final diagnosis.

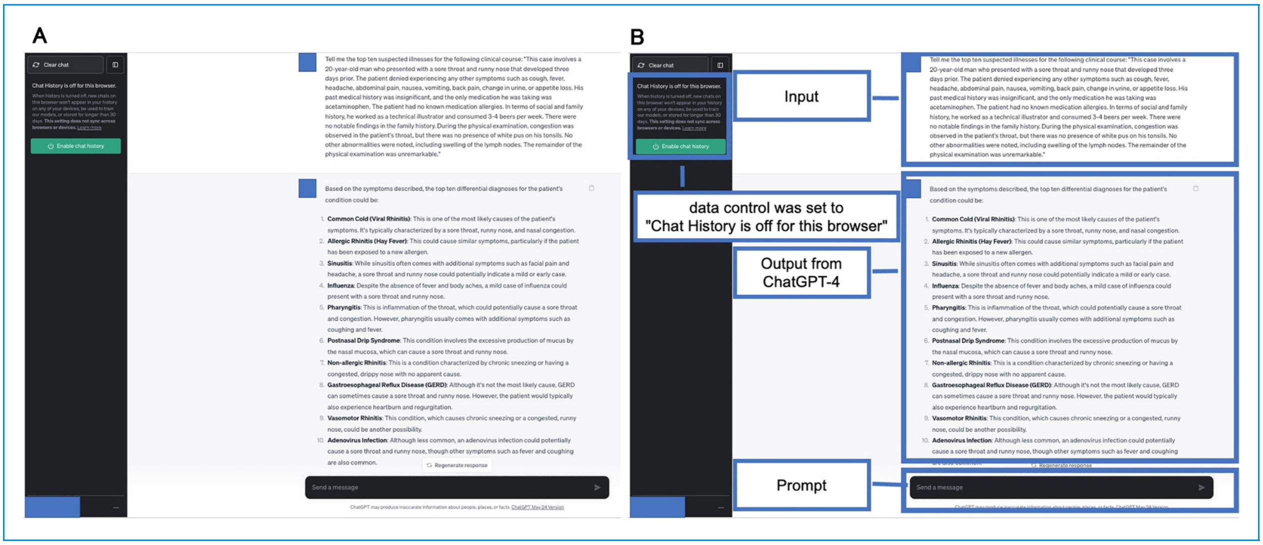
We also employed Google Gemini (previously Bard), an experimental AI application from Google, known for being updated daily. While it does not possess a specific version, Google Gemini is recognized for its robust language processing and generation skills. Figure 3A presents an example output of DDx list generated by Google Gemini for a sample case description. Figure 3B explains Figure 3A. Supplementary Table 2 lists the DDx generated by Google Gemini and the final diagnosis.

Additionally, LLaMA2 chatbot, an open-source application based on the LLaMA2 model, was available in versions 7B, 13B, and 70B. We employed the 70B version, the largest and most complex model among the options. The application allowed adjustments in settings such as temperature, top P, max sequence length, and prompt setting. The temperature parameter, which influences the model's randomness and creativity, ranged from 0.01 to 5.00. The top P parameter option adjusts the randomness of the model's predictions, ranging from 0.01 to 1.00. The max sequence length parameter specifies the maximum

Table 1. Details of generative artificial intelligences employed in this study.

	Generative artificial intelligences		
	ChatGPT-4	Google Gemini (previously Bard)	LLaMA2 chatbot
URL	https://chat.openai.com/	https://gemini.google.com/app	https://llama2.ai/
Company	OpenAI	Google	a16z (LLaMA2 model by Meta AI)
Version	March 24 version (default version, not code interpreter as beta feature)	No specific version	70B version
Access date	From 22 June to 29 June 2023	From 27 July to 1 August 2023	From 3 August to 8 August 2023
Adjustable settings	No setting (at that time, not available for custom instructions)	No setting	Temperature: 2.49, top P: 0.50, and max sequence length: 2048, default prompt setting
Data control	Chat history is off for this browser	Gemini activity is off	No data control
Fee	Fee-based	Free after users' login	Free after users' login
Knowledge cut-offs	September 2021	No cut-off date, due to constantly update	The pretraining data have a cut-off of September 2022, but some tuning data is more recent, up to July 2023
Others	Availability limitation: a cap of 50 messages every 3 h	The date was accessed before the system transitioned to Google Gemini	

ChatGPT-4: fourth-generation chat generative pre-trained transformer; LLaMA2: large language models by Meta AI 2.


A
B
Input
data control was set to "Chat History is off for this browser"
Output from ChatGPT-4
Prompt

Clear chat

Chat History is off for this browser.

When history is turned off, new chats on this browser won't appear in your history on any of your devices, be used to train our models, or stored for longer than 30 days. **This setting does not sync across browsers or devices.** Learn more

Enable chat history

Tell me the top ten suspected illnesses for the following clinical course: "This case involves a 20-year-old man who presented with a sore throat and runny nose that developed three days prior. The patient denied experiencing any other symptoms such as cough, fever, headache, abdominal pain, nausea, vomiting, back pain, change in urine, or appetite loss. His past medical history was insignificant, and the only medication he was taking was acetaminophen. The patient had no known medication allergies. In terms of social and family history, he worked as a technical illustrator and consumed 3-4 beers per week. There were no notable findings in the family history. During the physical examination, congestion was observed in the patient's throat, but there was no presence of white pus on his tonsils. No other abnormalities were noted, including swelling of the lymph nodes. The remainder of the physical examination was unremarkable."

Based on the symptoms described, the top ten differential diagnoses for the patient's condition could be:

1. **Common Cold (Viral Rhinitis)**: This is one of the most likely causes of the patient's symptoms. It's typically characterized by a sore throat, runny nose, and nasal congestion.
2. **Allergic Rhinitis (Hay Fever)**: This could cause similar symptoms, particularly if the patient has been exposed to a new allergen.
3. **Sinusitis**: While sinusitis often comes with additional symptoms such as facial pain and headache, a sore throat and runny nose could potentially indicate a mild or early case.
4. **Influenza**: Despite the absence of fever and body aches, a mild case of influenza could present with a sore throat and runny nose.
5. **Pharyngitis**: This is inflammation of the throat, which could potentially cause a sore throat and congestion. However, pharyngitis usually comes with additional symptoms such as coughing and fever.
6. **Postnasal Drip Syndrome**: This condition involves the excessive production of mucus by the nasal mucosa, which can cause a sore throat and runny nose.
7. **Non-allergic Rhinitis**: This is a condition characterized by chronic sneezing or having a congested, drippy nose with no apparent cause.
8. **Gastroesophageal Reflux Disease (GERD)**: Although it's not the most likely cause, GERD can sometimes cause a sore throat and runny nose. However, the patient would typically also experience heartburn and regurgitation.
9. **Vasomotor Rhinitis**: This condition, which causes chronic sneezing or a congested, runny nose, could be another possibility.
10. **Adenovirus Infection**: Although less common, an adenovirus infection could potentially cause a sore throat and runny nose, though other symptoms such as fever and coughing are also common.

Regenerate response

Send a message

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Clear chat

Chat History is off for this browser.

When history is turned off, new chats on this browser won't appear in your history on any of your devices, be used to train our models, or stored for longer than 30 days. **This setting does not sync across browsers or devices.** Learn more

Enable chat history

Tell me the top ten suspected illnesses for the following clinical course: "This case involves a 20-year-old man who presented with a sore throat and runny nose that developed three days prior. The patient denied experiencing any other symptoms such as cough, fever, headache, abdominal pain, nausea, vomiting, back pain, change in urine, or appetite loss. His past medical history was insignificant, and the only medication he was taking was acetaminophen. The patient had no known medication allergies. In terms of social and family history, he worked as a technical illustrator and consumed 3-4 beers per week. There were no notable findings in the family history. During the physical examination, congestion was observed in the patient's throat, but there was no presence of white pus on his tonsils. No other abnormalities were noted, including swelling of the lymph nodes. The remainder of the physical examination was unremarkable."

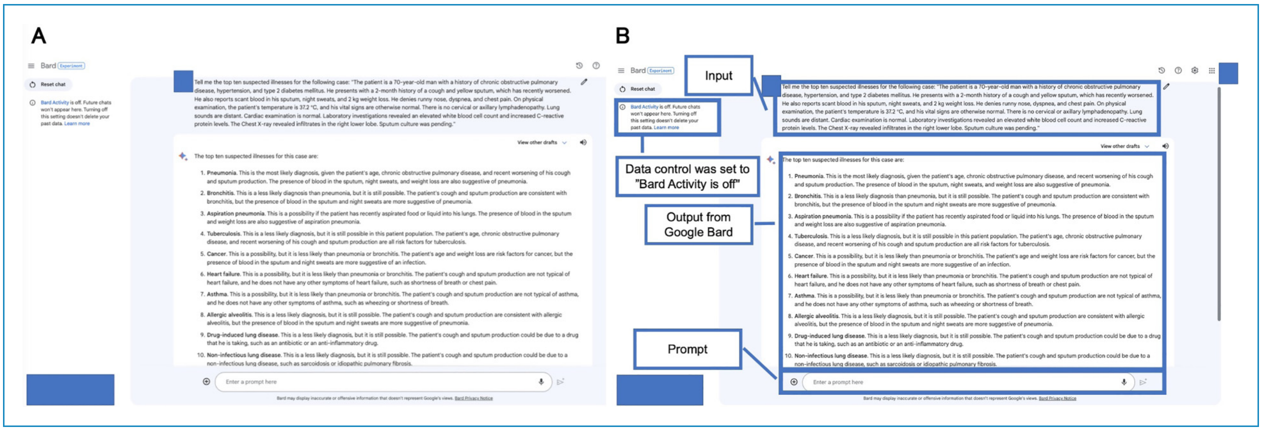
Based on the symptoms described, the top ten differential diagnoses for the patient's condition could be:

1. **Common Cold (Viral Rhinitis)**: This is one of the most likely causes of the patient's symptoms. It's typically characterized by a sore throat, runny nose, and nasal congestion.
2. **Allergic Rhinitis (Hay Fever)**: This could cause similar symptoms, particularly if the patient has been exposed to a new allergen.
3. **Sinusitis**: While sinusitis often comes with additional symptoms such as facial pain and headache, a sore throat and runny nose could potentially indicate a mild or early case.
4. **Influenza**: Despite the absence of fever and body aches, a mild case of influenza could present with a sore throat and runny nose.
5. **Pharyngitis**: This is inflammation of the throat, which could potentially cause a sore throat and congestion. However, pharyngitis usually comes with additional symptoms such as coughing and fever.
6. **Postnasal Drip Syndrome**: This condition involves the excessive production of mucus by the nasal mucosa, which can cause a sore throat and runny nose.
7. **Non-allergic Rhinitis**: This is a condition characterized by chronic sneezing or having a congested, drippy nose with no apparent cause.
8. **Gastroesophageal Reflux Disease (GERD)**: Although it's not the most likely cause, GERD can sometimes cause a sore throat and runny nose. However, the patient would typically also experience heartburn and regurgitation.
9. **Vasomotor Rhinitis**: This condition, which causes chronic sneezing or a congested, runny nose, could be another possibility.
10. **Adenovirus Infection**: Although less common, an adenovirus infection could potentially cause a sore throat and runny nose, though other symptoms such as fever and coughing

Regenerate response

Send a message

Figure 2. Example output from the fourth-generation chat generative pre-trained transformer (ChatGPT-4) for a sample case description (A) and the explanation (B).



Bard Experiment

Reset chat

Bard Activity is off. Future chats won't appear here. Turning off this setting doesn't delete your past data. Learn more

Tell me the top ten suspected illnesses for the following case: "The patient is a 70-year-old man with a history of chronic obstructive pulmonary disease, hypertension, and type 2 diabetes mellitus. He presents with a 2-month history of a cough and yellow sputum, which has recently worsened. He also reports scant blood in his sputum, night sweats, and 2 kg weight loss. He denies runny nose, dyspnea, and chest pain. On physical examination, the patient's temperature is 37.2 °C, and his vital signs are otherwise normal. There is no cervical or axillary lymphadenopathy. Lung sounds are distant. Cardiac examination is normal. Laboratory investigations revealed an elevated white blood cell count and increased C-reactive protein levels. The Chest X-ray revealed infiltrates in the right lower lobe. Sputum culture was pending."

View other drafts

The top ten suspected illnesses for this case are:

1. **Pneumonia**. This is the most likely diagnosis, given the patient's age, chronic obstructive pulmonary disease, and recent worsening of his cough and sputum production. The presence of blood in the sputum, night sweats, and weight loss are also suggestive of pneumonia.
2. **Bronchitis**. This is a less likely diagnosis than pneumonia, but it is still possible. The patient's cough and sputum production are consistent with bronchitis, but the presence of blood in the sputum and night sweats are more suggestive of pneumonia.
3. **Aspiration pneumonia**. This is a possibility if the patient has recently aspirated food or liquid into his lungs. The presence of blood in the sputum and weight loss are also suggestive of aspiration pneumonia.
4. **Tuberculosis**. This is a less likely diagnosis, but it is still possible in this patient population. The patient's age, chronic obstructive pulmonary disease, and recent worsening of his cough and sputum production are all risk factors for tuberculosis.
5. **Cancer**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's age and weight loss are risk factors for cancer, but the presence of blood in the sputum and night sweats are more suggestive of an infection.
6. **Heart failure**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's cough and sputum production are not typical of heart failure, and he does not have any other symptoms of heart failure, such as shortness of breath or chest pain.
7. **Asthma**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's cough and sputum production are not typical of asthma, and he does not have any other symptoms of asthma, such as wheezing or shortness of breath.
8. **Allergic alveolitis**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production are consistent with allergic alveolitis, but the presence of blood in the sputum and night sweats are more suggestive of pneumonia.
9. **Drug-induced lung disease**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production could be due to a drug that he is taking, such as an antibiotic or an anti-inflammatory drug.
10. **Non-infectious lung disease**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production could be due to a non-infectious lung disease, such as sarcoidosis or idiopathic pulmonary fibrosis.

Enter a prompt here

Bard may display inaccurate or offensive information that doesn't represent Google's views. Bard Privacy Notice

Bard Experiment

Reset chat

Bard Activity is off. Future chats won't appear here. Turning off this setting doesn't delete your past data. Learn more

Tell me the top ten suspected illnesses for the following case: "The patient is a 70-year-old man with a history of chronic obstructive pulmonary disease, hypertension, and type 2 diabetes mellitus. He presents with a 2-month history of a cough and yellow sputum, which has recently worsened. He also reports scant blood in his sputum, night sweats, and 2 kg weight loss. He denies runny nose, dyspnea, and chest pain. On physical examination, the patient's temperature is 37.2 °C, and his vital signs are otherwise normal. There is no cervical or axillary lymphadenopathy. Lung sounds are distant. Cardiac examination is normal. Laboratory investigations revealed an elevated white blood cell count and increased C-reactive protein levels. The Chest X-ray revealed infiltrates in the right lower lobe. Sputum culture was pending."

View other drafts

The top ten suspected illnesses for this case are:

1. **Pneumonia**. This is the most likely diagnosis, given the patient's age, chronic obstructive pulmonary disease, and recent worsening of his cough and sputum production. The presence of blood in the sputum, night sweats, and weight loss are also suggestive of pneumonia.
2. **Bronchitis**. This is a less likely diagnosis than pneumonia, but it is still possible. The patient's cough and sputum production are consistent with bronchitis, but the presence of blood in the sputum and night sweats are more suggestive of pneumonia.
3. **Aspiration pneumonia**. This is a possibility if the patient has recently aspirated food or liquid into his lungs. The presence of blood in the sputum and weight loss are also suggestive of aspiration pneumonia.
4. **Tuberculosis**. This is a less likely diagnosis, but it is still possible in this patient population. The patient's age, chronic obstructive pulmonary disease, and recent worsening of his cough and sputum production are all risk factors for tuberculosis.
5. **Cancer**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's age and weight loss are risk factors for cancer, but the presence of blood in the sputum and night sweats are more suggestive of an infection.
6. **Heart failure**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's cough and sputum production are not typical of heart failure, and he does not have any other symptoms of heart failure, such as shortness of breath or chest pain.
7. **Asthma**. This is a possibility, but it is less likely than pneumonia or bronchitis. The patient's cough and sputum production are not typical of asthma, and he does not have any other symptoms of asthma, such as wheezing or shortness of breath.
8. **Allergic alveolitis**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production are consistent with allergic alveolitis, but the presence of blood in the sputum and night sweats are more suggestive of pneumonia.
9. **Drug-induced lung disease**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production could be due to a drug that he is taking, such as an antibiotic or an anti-inflammatory drug.
10. **Non-infectious lung disease**. This is a less likely diagnosis, but it is still possible. The patient's cough and sputum production could be due to a non-infectious lung disease, such as sarcoidosis or idiopathic pulmonary fibrosis.

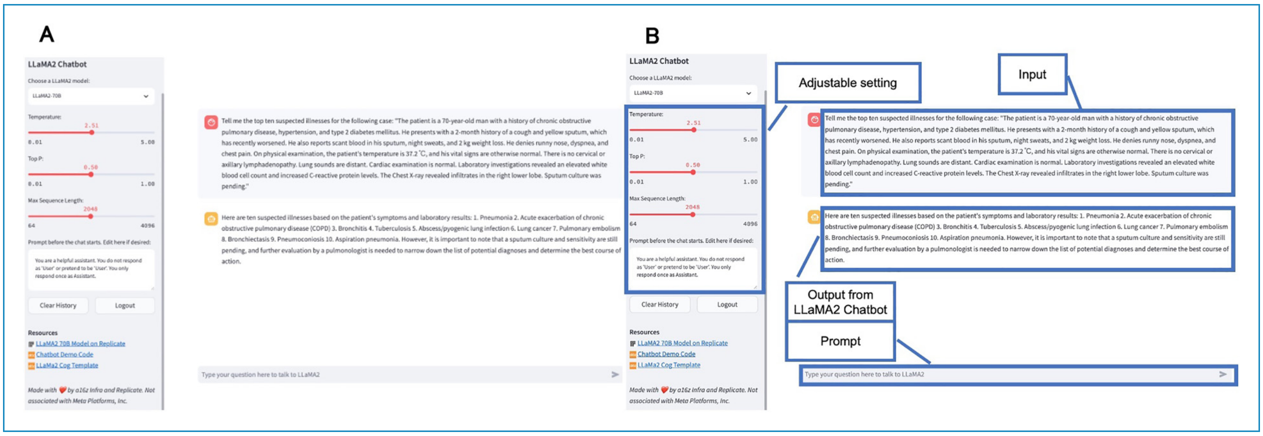
Enter a prompt here

Bard may display inaccurate or offensive information that doesn't represent Google's views. Bard Privacy Notice

Figure 3. Example output from Google Gemini (previously Bard) for a sample case description (A) and the explanation (B). The date was accessed before the system transitioned to Google Gemini.

number of word segments, tokens that can be generated by the model, ranging from 64 to 4096. The prompt setting was configured according to the role defined in the LLaMA2 model. Parameters set included a temperature of 2.49, a top *P* value of 0.50, and a max sequence length of 2048. We used the default setting for the following prompt: "You are a helpful assistant. You do not respond as 'User' or pretend to be 'User'. You only respond once as Assistant." The parameters were set based on preliminary testing, where various configurations were evaluated to determine the most effective settings for generating accurate and relevant DDx. Figure 4A illustrates an example output of DDx list generated by LLaMA2 chatbot for a sample case description. Figure 4B shows the explanation of Figure 4A. Supplementary Table 3 displays the DDx lists generated by LLaMA2 chatbot alongside the final diagnosis.

For all generative AIs, TH utilized a consistent prompt for each case description: "Tell me the top 10 suspected illnesses for the following case: (copy and paste each case description)." This prompt aimed to encourage the generative AIs to output a DDx list. The selection of the prompt was based on preliminary testing. The initial lists generated by the AIs served as the DDx lists. We cleared previous chat histories before inputting new case descriptions. To preclude any interference from previous interactions, we ensured that ChatGPT-4 and Google Gemini did not retain chat histories. Similarly, the LLaMA2 chatbot was used without a data control setting, preventing carryover effects from prior sessions.



LLaMA2 Chatbot

Choose a LLaMA2 model: LLaMA2-70B

Temperature: 2.51 (0.01–5.00)

Top P: 0.50 (0.01–1.00)

Max Sequence Length: 2048 (64–4096)

Prompt before the chat starts. Edit here if desired:

You are a helpful assistant. You do not respond as 'User' or pretend to be 'User'. You only respond once as Assistant.

Clear History | Logout

Resources
- LLaMA2 70B Model on Replicate
- Chatbot Demo Code
- LLaMa2 Cog Template

Made with ❤ by a16z Infra and Replicate. Not associated with Meta Platforms, Inc.

Tell me the top ten suspected illnesses for the following case: "The patient is a 70-year-old man with a history of chronic obstructive pulmonary disease, hypertension, and type 2 diabetes mellitus. He presents with a 2-month history of a cough and yellow sputum, which has recently worsened. He also reports scant blood in his sputum, night sweats, and 2 kg weight loss. He denies runny nose, dyspnea, and chest pain. On physical examination, the patient's temperature is 37.2 °C, and his vital signs are otherwise normal. There is no cervical or axillary lymphadenopathy. Lung sounds are distant. Cardiac examination is normal. Laboratory investigations revealed an elevated white blood cell count and increased C-reactive protein levels. The Chest X-ray revealed infiltrates in the right lower lobe. Sputum culture was pending."

Here are ten suspected illnesses based on the patient's symptoms and laboratory results: 1. Pneumonia 2. Acute exacerbation of chronic obstructive pulmonary disease (COPD) 3. Bronchitis 4. Tuberculosis 5. Abscess/pyogenic lung infection 6. Lung cancer 7. Pulmonary embolism 8. Bronchiectasis 9. Pneumoconiosis 10. Aspiration pneumonia. However, it is important to note that a sputum culture and sensitivity are still pending, and further evaluation by a pulmonologist is needed to narrow down the list of potential diagnoses and determine the best course of action.

Type your question here to talk to LLaMA2

Figure 4. Example output from large language model Meta AI 2 (LLaMA2) chatbot for a sample case description (A) and the explanation (B).

Evaluating the differential-diagnosis lists

Two GIM expert physicians (KM and TS; Tetsu Sakamoto), working independently, reviewed each final diagnosis and the DDx lists generated by AIs. The physicians assigned a binary code to each item. An item was coded "1" if it either accurately matched the final diagnosis with an acceptable specificity or was sufficiently close to the final diagnosis such that proper treatment would be initiated without compromising patient safety. A "0" was assigned if the items were those that were significantly different from the final diagnosis, such that treatment based on the responses would not have been effective.[30] Discrepancies between reviewers were resolved through consultation with an additional GIM expert physician (KT). All physicians in this evaluation process were blind to the specific generative AI that produced each list.

Measurement

Our primary outcome was the rate of the final diagnosis within the top 10 DDx lists generated by the AIs. The rate of the final diagnosis within the top 5 DDx lists and the rate of the final diagnosis as a top diagnosis were measured as secondary outcomes. Additionally, interrater reliability between the physicians' evaluation for the DDx lists was calculated as Cohen's kappa coefficient.

Statistical analysis

We analyzed the data using R version 4.2.2 (The R Foundation for Statistical Computing, Vienna, Austria) with the stats library (version 4.2.2). Categorical or binary variables were presented as numbers (percentages), and comparisons were made using the chi-square test. Given the multiple comparisons performed in this study, we applied the Bonferroni correction to adjust the significance levels and control the familywise error rate,[31] It adjusts the alpha or significance level by dividing it by the number of comparisons (three generative AIs in this study: ChatGPT-4, Google Gemini, and LLaMA2 chatbot) being made thereby setting a more stringent threshold for individual tests to be considered significant. The Bonferroni-corrected significance level was defined as a P value < 0.016. The Cohen's kappa coefficient was calculated using the irr package in R. The strength of agreement was categorized based on Cohen's kappa reference levels, which provided a standardized way of evaluating the level of agreement. A kappa coefficient of less than 0.40 indicated poor agreement; 0.41–0.75 suggested fair-to-good agreement; and 0.75–1.00 denoted very good agreement.[32]

Results

Included case reports

A total of 392 case reports were included. Three generative AIs, including ChatGPT-4, Google Gemini, and LLaMA2 chatbot, answered all 392 case descriptions. Table 2 displays the representative DDx list by three generative AIs. The DDx lists for all case descriptions generated by ChatGPT-4, Google Gemini, and LLaMA2 chatbot are shown in Supplementary Tables 1–3, respectively.

Diagnostic performance

Table 3 and Figure 5 present the diagnostic accuracy of three generative AIs. Regarding the rate of final diagnoses within the top 10 DDx lists, ChatGPT-4 demonstrated 86.7% (340/392), which was statically higher than those of Google Gemini (269/392, 68.6%, $P < 0.001$) and LLaMA2 chatbot (214/392, 54.6%, $P < 0.001$). Similarly, the top diagnoses by ChatGPT-4 were 54.6% (214/392),

Table 2. Representative differential-diagnosis list by generative artificial intelligences and final diagnosis.

Number	Final diagnosis	ChatGPT-4	Google Gemini	LLaMA2 Chatbot
1	Central diabetes insipidus	1. Central diabetes insipidus 2. Nephrogenic diabetes insipidus (NDI) 3. Primary polydipsia (psychogenic polydipsia) 4. Hypercalcemia 5. Hyperthyroidism 6. Cushing's syndrome 7. Uncontrolled diabetes mellitus 8. Sickle cell disease 9. Kidney disease/failure 10. Liver disease	1. Central diabetes insipidus 2. Nephrogenic diabetes insipidus 3. Primary polydipsia 4. Psychogenic polydipsia 5. Situational polydipsia 6. Drug-induced polydipsia 7. Hypothyroidism 8. Hypercalcemia 9. Adrenal insufficiency 10. Diabetes mellitus	1. Diabetes insipidus 2. Central diabetes insipidus 3. Nephrogenic diabetes insipidus 4. Type 2 diabetes mellitus 5. Polycystic kidney disease 6. Post-streptococcal glomerulonephritis 7. Hypertension 8. Hyperglycemia 9. Polyuria 10. Polydipsia
2	Concurrent arterial infarct	1. Post-acute sequelae of SARS-CoV-2 infection (PASC) 2. Cerebral venous sinus thrombosis (CVST) 3. Acute disseminated encephalomyelitis 4. Neurological manifestation of sepsis or systemic infection 5. Metabolic encephalopathy (due to electrolyte imbalances, although many tests were normal) 6. Encephalitis (inflammatory brain disease, viral or otherwise) 7. Autoimmune encephalitis 8. Neurological manifestation of a systemic autoimmune disease (e.g., systemic lupus erythematosus, Sjögren's syndrome, vasculitis) despite normal autoimmune tests 9. Paraneoplastic syndromes (neurological disorders triggered by cancers) 10. Central nervous system vasculitis	1. Venous sinus thrombosis 2. Bilateral subcortical arterial infarcts 3. SARS-CoV-2 infection 4. Systemic lupus erythematosus 5. Antiphospholipid syndrome 6. Moyamoya disease 7. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) 8. Atrial fibrillation 9. Cardioembolism 10. Subarachnoid hemorrhage	1. Cerebral venous sinus thrombosis 2. Bilateral subcortical arterial infarcts 3. SARS-CoV-2 infection-related encephalopathy 4. Hypoxic-ischemic encephalopathy 5. Bilateral cerebral vasculitis 6. Moyamoya disease 7. Systemic lupus erythematosus 8. Sjögren's syndrome 9. Antineutrophil cytoplasmic antibody-associated vasculitis 10. Antiphospholipid syndrome
3	Levofloxacin-associated bullous pemphigoid	1. Bullous pemphigoid 2. Pemphigus vulgaris 3. Drug-induced eruptions 4. Stevens-Johnson syndrome or toxic epidermal necrolysis 5. Contact dermatitis or allergic dermatitis 6. Infectious causes 7. Systemic lupus	1. Disseminated herpes zoster 2. Acute cellular and antibody-mediated rejection 3. Acute interstitial nephritis 4. Drug reaction 5. Bullous pemphigoid 6. Staphylococcal scalded skin syndrome 7. Toxic epidermal necrolysis	1. Sepsis due to UTI 2. Pyelonephritis 3. Pneumonia 4. Hemolytic anemia 5. Shingles 6. BK nephropathy 7. Acute kidney injury 8. Hypersensitivity reaction

(continued)

Table 2. Continued.

Number	Final diagnosis	ChatGPT-4	Google Gemini	LLaMA2 Chatbot
		erythematosus (SLE) 8. Vasculitis 9. Sepsis or bacteremia 10. Pulmonary embolism or heart disease	8. Stevens–Johnson syndrome 9. Sepsis 10. Malignant hypertension	9. Bullous pemphigoid 10. Psoriasis

ChatGPT-4: fourth-generation chat generative pre-trained transformer; LLaMA2: large language models by Meta AI 2.

Table 3. Diagnostic accuracy of generative artificial intelligences.

				P value[a]		
	ChatGPT-4	Google Gemini	LLaMA2 Chatbot	ChatGPT-4 vs Google Gemini	ChatGPT-4 vs LLaMA2 Chatbot	Google Gemini vs LLaMA2 Chatbot
The rate of final diagnosis within the top 10 DDx lists, *n* (%)	340 (86.7)	269 (68.6)	214 (54.6)	< 0.001	< 0.001	<0.001
The rate of final diagnosis within the top five DDx lists, *n* (%)	307 (78.3)	235 (59.9)	177 (45.2)	< 0.001	< 0.001	< 0.001
The rate of final diagnosis as top diagnosis, *n* (%)	214 (54.6)	123 (31.4)	90 (23.0)	< 0.001	< 0.001	0.010

ChatGPT-4: fourth-generation chat generative pre-trained transformer; DDX: differential-diagnosis; LLaMA2: large language models by Meta AI 2.
[a]Chi-squared test.

which was statically higher than those of Google Gemini (123/392, 31.4%, $P < 0.001$) and LLaMA2 chatbot (90/392, 23.0%, $P < 0.001$). Moreover, Google Gemini was more accurate than LLaMA2 chatbot within the top 10 (269/392, 68.6% vs 214/392, 54.6%, respectively; $P < 0.001$) and as the top diagnosis (213/392, 31.4% vs 90/392, 23.0%, respectively; $P = 0.010$). We observed very good agreement among physicians' evaluations for the DDx lists, with concordance in 88.9% (1045/1176) of cases. The kappa coefficient was 0.76.

Discussion

Principal results

Several key findings emerged from this study, providing insights into the diagnostic performance of different generative AIs. ChatGPT-4 demonstrated superior diagnostic accuracy compared to Google Gemini and LLaMA2 chatbot. The primary distinction lies in the LLMs powering these AIs, yet this superiority could be attributed to ChatGPT's earlier introduction to the public, allowing more time for reinforcement. The timing of release may have provided more time for iterative improvements based on user feedback and real-world performance data.[33] Additionally, ChatGPT-4's fee-based application model could contribute to its superior performance. The fee-based model could influence the frequency of model updates and the extent to which user feedback is incorporated. The superiority of ChatGPT-4 highlighted that generative AIs have the potential of rapid improvement in diagnostic performance, even without specific training within the diagnostic domain.

In contrast to ChatGPT-4's marked success, our analysis revealed a difference in the performance of the other AI platforms tested. Google Gemini performed better diagnostically than LLaMA2 chatbot. This result indicates that LLaMA2 chatbot requires more refinement in the diagnostic domain. While the fundamental difference was the underlying LLMs of these AIs, the variability in performance could be due to the adjustable parameters unique to the LLaMA2 chatbot in the current study, an aspect of flexibility absent in ChatGPT-4 and Google Gemini. For example, LLaMA2 chatbot allows adjustments to parameters including temperature, which influences randomness of the generated responses. Therefore, LLaMA2

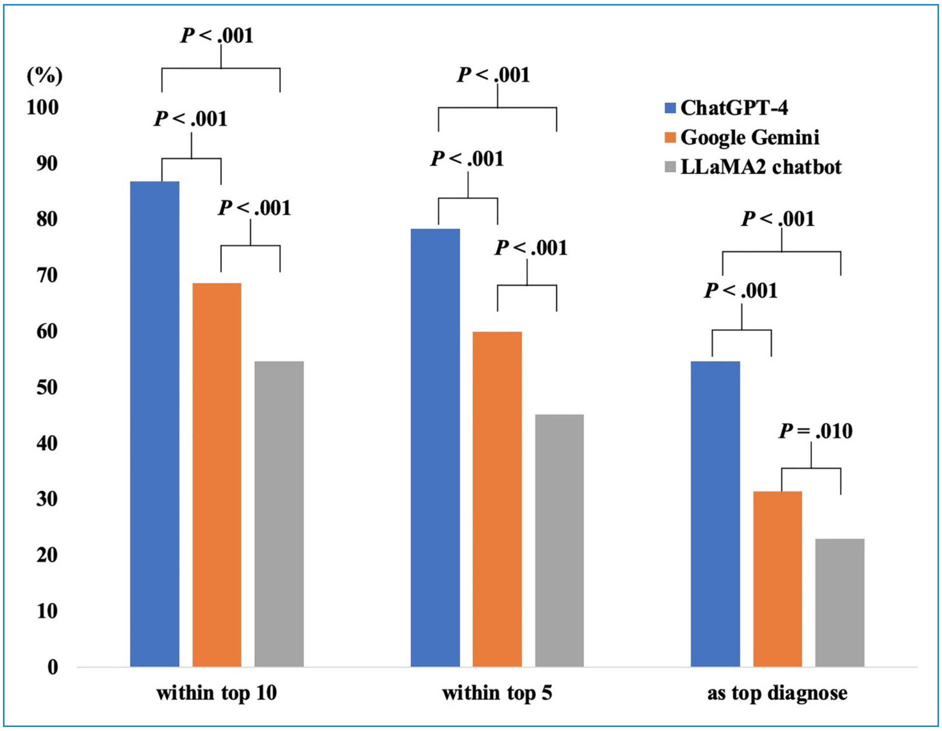


Figure 5. Diagnostic performance of generative artificial intelligences.

chatbot might improve with the optimization of these parameters.

We found that these generative AIs exhibited impressive diagnostic accuracy when presented with complex case descriptions, even without specialized medical training and tuning. These results highlight the necessity of recognizing the differences in diagnostic performance among generative AIs, especially for complex case series across comprehensive medical fields like GIM. Generative AIs could serve as a form of collective intelligence, pooling knowledge and insights from various sources, proficient in verifying medical diagnoses made by healthcare professionals.

The differences in diagnostic performance among ChatGPT-4, Google Gemini, and LLaMA2 chatbot can be attributed to the underlying technologies, training data, and iterative improvements specific to each AI. ChatGPT-4's early introduction, extensive training data, and commercial model likely contribute to its superior performance. In contrast, Google Gemini benefits from Google's advanced AI infrastructure, while LLaMA2's performance indicates the need for further refinement and parameter optimization. Understanding these factors is essential for developing more effective generative AI systems in the medical diagnostic domain.

Limitations

A primary limitation of this study is its reliance on case materials from a single-case report journal. While this ensured consistency in the data, it introduced selection bias, limiting the generalizability to a broader range of medical scenarios. Different journals may emphasize distinct aspects in their case descriptions. This diversity could impact the diagnostic performance of generative AIs. Therefore, our results may not fully represent the diversity of real-world clinical situations. It is imperative for healthcare professionals to evaluate the efficacy and safety of generative AI systems in real-world cases across various settings to ascertain their capabilities comprehensively.[34] Furthermore, we excluded pediatric cases and those primarily focused on management from our analysis. This exclusion may have influenced the results and introduced additional selection bias.

Additionally, we did not compare the diagnostic performance of physicians directly with generative AIs. This decision stemmed from the view that generative AIs are intended to supplement, not replace, physicians.[35] A comparative study could provide deeper insights into how generative AIs perform relative to medical professionals and highlight areas where AI can most effectively augment clinical decision making.

Moreover, our study focused solely on textual data, excluding tables and figures, which could be crucial for assessing the diagnostic performance of generative AIs. While our previous study revealed that visual information did not improve the diagnostic accuracy of ChatGPT-4 for case report series,[36] visual information often plays a significant role in medical diagnostics, and its exclusion may have limited the scope of our findings.

As for the limitations of generative AIs, it is crucial to note that these AI tools lack formal medical approval. Additionally, optimal prompt settings, such as temperature, Top P, max sequence length, the setting of prompt, and prompt text, remain to be established. Moreover, our study did not encompass all available generative AI platforms.[10] Future research should comprehensively explore these generative AIs.

Another interesting aspect that was not fully explored in the current study is the comparison of different AI systems' performance with regard to their training data. Different AI systems are trained on varied datasets, which can significantly influence their diagnostic capabilities and performance. Future research should delve into how the nature and diversity of training data affect the diagnostic accuracy and generalizability of these systems.

Lastly, the rapid evolution of AI technology means that our findings may quickly become outdated. The update from ChatGPT-4 to ChatGPT-4o, from Google Bard to Google Gemini,[13] and from LLaMA2 to LLaMA3 exemplifies the fast-paced advancements in this field. Ongoing research is necessary to keep up with these developments and to continually reassess the performance and safety of new generative AI iterations.

Comparison with prior work

Diagnostic performance of the DDx lists for ChatGPT-4 and Google Gemini is summarized in Table 4. Relative to our previous study[21] on the GIM case series, the differences in diagnostic accuracy by ChatGPT-4 were no more than 6% within the top 10 DDx list (340/392, 86.7% vs 43/52, 83%, respectively), within the top 5 DDx list (307/392, 78.3% vs 42/52, 81%, respectively), and as top diagnosis (214/392, 54.6% vs 31/52, 60%, respectively). In contrast, compared to another study[17] of clinicopathologic conference from *New England Journal of Medicine*, our findings showed higher diagnostic accuracy in the DDx lists generated by ChatGPT-4 (340/392, 86.7% vs 45/70, 64%, respectively) and its top diagnosis (214/392, 54.6% vs 27/70, 39%, respectively). This consistency with the former study was partly attributed to similarities in case difficulty, as both sets of cases presented complex diagnostic challenges. Additionally, the evaluation methods used in the current study closely mirrored those of the previous one, enhancing the comparability of results. Despite the complexity of cases sourced from various journals, the results highlighted the evolving capabilities of generative AIs. On the other hand, the variation observed when comparing with the latter study could partly be attributed to differences in study designs, case materials, and evaluation methods.

Regarding Google Gemini,[23] the present finding revealed higher diagnostic accuracy within the top 10 DDx list (269/392, 68.6% vs 22/52, 42.3%, respectively), within the top 5 DDx list (235/392, 59.9% vs 21/52, 40.4%, respectively), and as top diagnosis (123/392, 31.4% vs 15/52, 28.8%, respectively). This enhancement can be partly attributed to the continuous updates to the model. The timeframe for accessing the AI in the current study spanned from July to August, while the previous study was drawn in June.

In the context of radiological differential diagnoses, a study[37] found acceptance rates of 85% and 69% for the top five DDx generated by ChatGPT-3.5 and Google Bard (currently Gemini), respectively. Our findings for Google Gemini (60%) show a similar level. Additionally, our finding for the next-generation model, ChatGPT-4 (78%), indicates that while improvements are not significant, variations in case types and evaluation methods likely influence these results.

Table 4. Diagnostic performance of differential-diagnosis lists for ChatGPT-4 and Google Gemini.

	ChatGPT-4			Google Gemini	
	This study	GIM case reports [a]	Clinicopathologic conference from *New England Journal of Medicine* [b]	This study	GIM case reports [a]
The rate of final diagnosis within the top 10 DDx lists, *n*/*N* (%)	340/392 (86.7)	43/52 (82.7)	45/70 (64.3)	269/392 (68.6)	22/52 (42.3)
The rate of final diagnosis within the top five DDx lists, *n*/*N* (%)	307/392 (78.3)	42/52 (80.8)	NA	235/392 (59.9)	21/52 (40.4)
The rate of final diagnosis as top diagnosis, *n*/*N* (%)	214/392 (54.6)	31/52 (59.6)	27/70 (38.5)	123/392 (31.4)	15/52 (28.8)

ChatGPT-4: fourth-generation chat generative pre-trained transformer; DDX: differential-diagnosis; GIM: general internal medicine.

[a] JMIR Med Inform. 2023 Oct 9:11:e48808.

[b] JAMA. 2023;330(1):78–80.

Compared to the symptom checkers,[6] ChatGPT-4 showed higher rates of final diagnoses within the top 10 DDx lists (340/392, 86.7% vs 60.9–76.9%, respectively), while Google Gemini performed at a similar level (269/392, 68.6% vs 60.9–76.9%, respectively) and LLaMA2 chatbot lower (214/392, 54.6% vs 60.9–76.9%, respectively). Regarding the DDx generators,[38] ChatGPT-4 demonstrated higher rates (340/392, 86.7% vs 63–77%, respectively), while Google Gemini performed at the same level (269/392, 68.6% vs 63–77%, respectively) and LLaMA2 chatbot lower (214/392, 54.6% vs 63–77%, respectively). These findings underscore ChatGPT-4's potential as a reliable adjunct in medical diagnostics, while Google Gemini maintains comparable performance, and LLaMA2 chatbot lags behind.

Future directions

Based on this study, future research should explore real-world cases in various settings—such as different medical specializations, varying levels of healthcare, or diverse geographical locations—after securing the necessary regulatory approval for medical use. This approval, which includes clearance or authorization from health authorities, ensures that the generative AI systems adhere to patient safety and efficacy standards.

Additionally, research should investigate the fine tuning of AI models for medical applications and explore ensemble methods that integrate multiple generative AIs. Specifically, future studies should focus on establishing optimal prompt settings for these AIs, including parameters such as temperature, Top P, max sequence length, and prompt text. Understanding the impact of these settings on diagnostic accuracy is crucial for optimizing AI performance in clinical practice.

As the integration of technology and healthcare deepens, there is also a pressing need to evaluate the efficacy of collaborative efforts between healthcare professionals and generative AIs in clinical settings. Future research should investigate the potential benefits and challenges of human–AI collaboration in the diagnostic process, aiming to enhance the overall quality of care.

Moreover, fostering effective collaboration between generative AIs, healthcare professionals including GIM physicians, and patients warrants attention for a team-based approach going forward.[39] This includes exploring how these systems can be integrated into daily clinical practice, ensuring seamless interaction between AI tools and medical professionals. Additionally, regulatory considerations must be underlined to ensure that AI integration in clinical practice adheres to safety, ethical, and legal standards, thereby protecting patient welfare and maintaining trust in AI-assisted healthcare.

Conclusions

In summary, this study highlighted the diagnostic performance of generative AIs, including ChatGPT-4, Google Gemini, and LLaMA2 chatbot. ChatGPT-4 exhibited higher diagnostic accuracy than the other platforms across comprehensive medical fields, like GIM. These findings emphasize the importance of understanding the differences in diagnostic performance among generative AIs, especially in complex case series across comprehensive medical fields. Additionally, generative AIs, especially those exhibiting superior accuracy like ChatGPT-4, possess the potential to be pivotal diagnostic support tools in the near future. However, the variability in performance among the different AIs tested indicates the necessity for continued refinement and validation of these tools, especially in real-world clinical contexts.

Acknowledgments: This study was made possible using the resources from the Department of Diagnostic and Generalist Medicine, Dokkyo Medical University.

Contributorship: TH and YH researched literature and conceived the study. TH was involved in protocol development, gaining ethical approval, participant recruitment, and data analysis. TH wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Not applicable.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by JSPS KAKENHI [Grant Number 22K10421].

Guarantor: TH

Informed consent: Not applicable for studies not involving humans.

Trial Registration: Not applicable.

ORCID iD: Takanobu Hirosawa https://orcid.org/0000-0002-3573-8203

Supplemental material: Supplemental material for this article is available online.

References

1. Yang D, Fineberg HV and Cosby K. Diagnostic excellence. *JAMA* 2021; 326: 1905–1906.
2. Watari T and Schiff GD. Diagnostic excellence in primary care. *J Gen Fam Med* 2023; 24: 143–145.

3. Staal J, Hooftman J, Gunput STG, et al. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. *BMJ Qual Saf* 2022; 31: 899–910.
4. Singh H, Connor DM and Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. *Br Med J* 2022; 376: e068044.
5. Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med* 2020; 3: 17.
6. Schmieding ML, Kopka M, Schmidt K, et al. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. *J Med Internet Res* 2022; 24: e31810.
7. Bond WF, Schwartz LM, Weaver KR, et al. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med* 2012; 27: 213–219.
8. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, et al. Diagnostic accuracy in family medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. *Diagnosis (Berl)* 2018; 5: 71–76.
9. Jain B. The key role of differential diagnosis in diagnosis. *Diagnosis (Berl)* 2017; 4: 239–240.
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023; 29: 1930–1940.
11. Thirunavukarasu AJ, Mahmood S, Malem A, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. *PLOS Digit Health*. 2024; 3. DOI: 10.1371/journal.pdig.0000341.
12. OpenAI. GPT-4 Technical Report2023 01 March 2023:[arXiv:2303.08774 p.].
13. Patrizio A. Google Gemini: TechTarget; 2023 Available from: https://www.techtarget.com/searchenterpriseai/definition/Google-Gemini.
14. Anil R, Dai AM, Firat O, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023.
15. Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
16. Haug CJ and Drazen JM. Artificial intelligence and machine learning in clinical medicine. *N Engl J Med* 2023; 388: 1201–1208.
17. Kanjee Z, Crowe B and Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023; 330: 78–80.
18. Liu J, Wang C and Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023; 25: e48568.
19. Kung T, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2: e0000198.
20. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Gemini on a Neurosurgery Oral Boards Preparation Question Bank. *medRxiv*. 2023:2023.04. 06.23288265.
21. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med Inform* 2023; 11: e48808.
22. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT versus Google Gemini. *Radiology* 2023; 307: e230922.
23. Hirosawa T, Mizuta K, Harada Y, et al. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. *Am J Med* 2023; 136: 1119–1123.
24. Takagi S, Watari T, Erabi A, et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023; 9: e48002.
25. Detsky AS, Gauthier SR and Fuchs VR. Specialization in medicine: how much is appropriate? *JAMA* 2012; 307: 463–464.
26. Graber ML and Mathew A. Performance of a web-based clinical diagnosis support system for internists. *J Gen Intern Med* 2008; 23: 37–40.
27. Ohara N, Takada T, Seki Y, et al. A 75-year-old woman with a 5-year history of controlled type 2 diabetes Mellitus presenting with polydipsia and polyuria and a diagnosis of central diabetes insipidus. *Am J Case Rep* 2022; 23: e938482.
28. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or Potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. *JMIR Med Educ.* 2024; 10: e51148.
29. Funk PF, Hoch CC, Knoedler S, et al. ChatGPT's response consistency: a study on repeated queries of medical examination questions. *Eur J Investig Health Psychol Educ* 2024; 14: 657–668.
30. Krupat E, Wormwood J, Schwartzstein RM, et al. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. *Med Educ* 2017; 51: 1127–1137.
31. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 2014; 34: 502–508.
32. Fleiss JL, Levin B and Paik MC. *Statistical methods for rates and proportions*. New York: John Wiley & Sons, 2003.
33. Bajwa J, Munir U, Nori A, et al. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J* 2021; 8: e188–ee94.
34. Painter A, Hayhoe B, Riboli-Sasco E, et al. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. *J Med Internet Res* 2022; 24: e37408.
35. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *Digit Health* 2023; 9: 20552076231186520.
36. Hirosawa T, Harada Y, Tokumasu K, et al. Evaluating ChatGPT-4's diagnostic accuracy: impact of visual data integration. *JMIR Med Inform* 2024; 12: e55627.
37. Sarangi PK, Irodi A, Panda S, et al. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. *Indian J Radiol Imaging* 2024; 34: 269–275.
38. Riches N, Panagioti M, Alam R, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. *PLoS One* 2016; 11: e0148991.
39. Harris E. Large language models answer medical questions accurately, but can't match Clinicians' knowledge. *JAMA* 2023; 330: 792–794.