Original Paper

# Comparative Analysis of Diagnostic Performance: Differential Diagnosis Lists by LLaMA3 Versus LLaMA2 for Case Reports

Takanobu Hirosawa[1], MD, PhD; Yukinori Harada[1], MD, PhD; Kazuki Tokumasu[2], MD, PhD; Tatsuya Shiraishi[3,4], MD; Tomoharu Suzuki[5], MD; Taro Shimizu[1], MSc, MPH, MBA, MD, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Shimotsuga, Japan

[2]Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

[3]Higashinihonbashinaika Clinic, Tokyo, Japan

[4]Ubie, Inc, Tokyo, Japan

[5]Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan

**Corresponding Author:**
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga, 321-0293
Japan
Phone: 81 0282861111
Email: hirosawa@dokkyomed.ac.jp

## Abstract

**Background:** Generative artificial intelligence (AI), particularly in the form of large language models, has rapidly developed. The LLaMA series are popular and recently updated from LLaMA2 to LLaMA3. However, the impacts of the update on diagnostic performance have not been well documented.

**Objective:** We conducted a comparative evaluation of the diagnostic performance in differential diagnosis lists generated by LLaMA3 and LLaMA2 for case reports.

**Methods:** We analyzed case reports published in the *American Journal of Case Reports* from 2022 to 2023. After excluding nondiagnostic and pediatric cases, we input the remaining cases into LLaMA3 and LLaMA2 using the same prompt and the same adjustable parameters. Diagnostic performance was defined by whether the differential diagnosis lists included the final diagnosis. Multiple physicians independently evaluated whether the final diagnosis was included in the top 10 differentials generated by LLaMA3 and LLaMA2.

**Results:** In our comparative evaluation of the diagnostic performance between LLaMA3 and LLaMA2, we analyzed differential diagnosis lists for 392 case reports. The final diagnosis was included in the top 10 differentials generated by LLaMA3 in 79.6% (312/392) of the cases, compared to 49.7% (195/392) for LLaMA2, indicating a statistically significant improvement ($P$<.001). Additionally, LLaMA3 showed higher performance in including the final diagnosis in the top 5 differentials, observed in 63% (247/392) of cases, compared to LLaMA2's 38% (149/392, $P$<.001). Furthermore, the top diagnosis was accurately identified by LLaMA3 in 33.9% (133/392) of cases, significantly higher than the 22.7% (89/392) achieved by LLaMA2 ($P$<.001). The analysis across various medical specialties revealed variations in diagnostic performance with LLaMA3 consistently outperforming LLaMA2.

**Conclusions:** The results reveal that the LLaMA3 model significantly outperforms LLaMA2 per diagnostic performance, with a higher percentage of case reports having the final diagnosis listed within the top 10, top 5, and as the top diagnosis. Overall diagnostic performance improved almost 1.5 times from LLaMA2 to LLaMA3. These findings support the rapid development and continuous refinement of generative AI systems to enhance diagnostic processes in medicine. However, these findings should be carefully interpreted for clinical application, as generative AI, including the LLaMA series, has not been approved for medical applications such as AI-enhanced diagnostics.

***(JMIR Form Res 2024;8:e64844)*** doi: 10.2196/64844

**KEYWORDS**

artificial intelligence; clinical decision support system; generative artificial intelligence; large language models; natural language processing; NLP; AI; clinical decision making; decision support; decision making; LLM: diagnostic; case report; diagnosis; generative AI; LLaMA

# Introduction

## Artificial Intelligence in Medicine

The concept of artificial intelligence (AI) dates back to the 1950s when the potential for machines to mimic human intelligence first began to be explored [1]. Since then, AI technologies, particularly in areas such as neural networks, natural language processing (NLP), and large language models (LLMs), have advanced substantially. These advancements have been driven by significant computational developments and the vast data available in the digital world. Recently, access to these technologies has also become more straightforward, requiring less specific knowledge and fewer resources.

In the realm of AI, neural networks form a foundational concept. These networks mimic the complex interconnections of neurons in the human brain, featuring synapse-like connections that facilitate dynamic learning and adoption. Unlike traditional technologies that rely on static algorithms, neural networks are designed to iteratively adjust the connections between nodes [2]. NLP enables computers to understand and process human language, facilitating tasks such as text translation, voice command response, and data extraction from complex sources. LLMs, advanced forms of NLP, train on extensive corpora of text to generate coherent and contextually relevant text [3]. These technologies have enabled complex models to achieve improved performance and address challenges that traditional approaches cannot handle, such as analyzing large volumes of data to identify patterns that may not be visible to human analysts.

These advancements are now widespread across various sectors, notably in the medical field. Generative AI systems, such as the GPT series developed by OpenAI, Google's Gemini, and LLaMA, have demonstrated considerable value in research, education, and potential future clinical applications [4,5]. They have the potential to support medical professionals, patients, and their families, by aiding them in making informed clinical decisions based on comprehensive data analysis.

## Generative AI in Medicine

In the medical field, generative AI has been pivotal in advancing diagnostic processes, developing treatment protocols, enabling personalized medicine, and managing patient care [6]. By analyzing vast datasets, generative AI uncovers patterns not immediately obvious to medical professionals, providing crucial insights that lead to improved patient outcomes. For example, generative AI systems are instrumental in enhancing clinical decision-making, optimizing clinical workflows, and improving patient outcomes [7]. Specifically, in diagnosis, generative AI enhances the medical interview process by visualizing the patient's perspective [8], expands the scope of differential diagnosis lists, and supports clinical reasoning [9,10].

## From LLaMA2 to LLaMA3

The evolution of generative AI systems has been notably rapid, primarily due to their ability to integrate user feedback and continuously update from expanded datasets. This iterative improvement is evident in the progression from GPT-3 to GPT-4, and more recently to GPT-4o and OpenAI o1 [11,12]. Similarly, other systems such as Bard have evolved into more advanced versions such as Gemini and Gemini Advanced [13]. In this dynamic landscape, the LLaMA series has also undergone upgrades, moving from LLaMA2 to LLaMA3, enhancing their capabilities [14].

## Generative AI in Diagnostics

In diagnostics, generative AI systems have the potential to enhance diagnostic performance. These systems excel at processing and interpreting complex clinical data from diverse courses such as electronic health records, imaging studies, and genomic data. Notably, the GPT series has demonstrated considerable diagnostic performance in medical benchmarks and complex case analyses [15]. While significant strides have been made, studies have indicated that other LLM models, such as LLaMA2, require substantial refinement for optimal application in diagnostics [16,17]. Our own study revealed that the diagnostic performance by LLaMA2 was inferior to those of ChatGPT-4 and Gemini for case-report series [18]. This necessitates ongoing development to improve model accuracy and reliability, ensuring they meet clinical standards and effectively support diagnostic decision-making.

## Study Aims

Despite these advancements, the diagnostic capabilities of updated AI models such as LLaMA3 have not been comprehensively explored. There is a particular lack of comparative studies examining the improvements in diagnostic performance from LLaMA2 to LLaMA3. In this context, our study aims to fill this gap by assessing and comparing the diagnostic performance of LLaMA3 to LLaMA2. Specifically, we intend to evaluate their effectiveness in generating differential diagnosis lists for comprehensive case reports. This comparison will explain the evolutionary benefits of the generative AI system upgrade and their practical implications in future diagnostics.

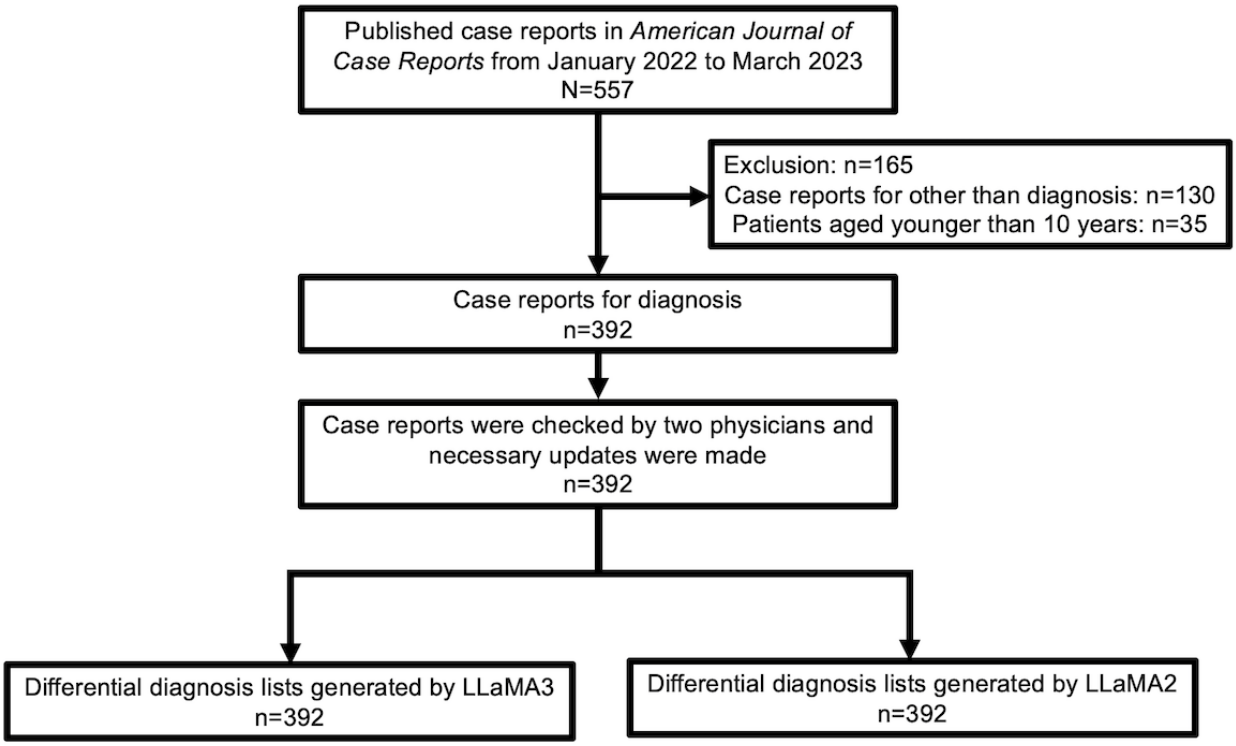
# Methods

## Overview

This was an experimental study using publicly available generative AI systems and published case reports. The entire study was conducted at the Department of Diagnostic and Generalist Medicine (General Internal Medicine), Dokkyo Medical University, Japan. This study consisted of four components, including preparing case reports, generating differentials by AIs, evaluating the differentials, and analysis.

The flowchart, including preparing case reports and generating differentials, is shown in Figure 1.

**Figure 1.** The flowchart, including preparing case reports and generating differentials.




## Ethical Considerations

We used published case reports; therefore, ethical approval was inapplicable.

## Case Reports

We used the dataset from our previous research [18]. Our inclusion criteria included case reports published in the *American Journal of Case Reports* from January 2022 to March 2023. We excluded nondiagnostic cases and pediatric cases. These exclusion criteria were adopted from a previous study for a clinical decision support system [19]. For the included case reports, we refined the text data for input. This process involved extracting the clinical narrative from each case report up until the stated final diagnosis. We carefully removed any sections that included clinical assessments or subjective interpretations by the authors to minimize the risk of biasing the AI's output. This editing was designed to ensure that the input to the AI models was focused on clinical information essential for generating accurate differential diagnoses. The final diagnoses were typically written by the authors. The main investigator, TH, conducted this process, which was validated by another coinvestigator, YH. Details of preparing case reports are shown in Multimedia Appendix 1.

## Differentials Generated by AIs

We used popular generative AI systems developed by Meta AI, LLaMA3 and LLaMA2, to generate differentials. LLaMA3 offers 8B and 70B versions, while LLaMA2 includes 7B, 13B, and 70B versions. For our study, we used the most capable models, the 70B versions. The main investigator, TH, inputted the same cases into both LLaMA3 and LLaMA2 using the same prompt to generate the top 10 differential diagnosis lists.

Both LLaMA3 and LLaMA2 allowed for several adjustable settings to control the output, including temperature, top-P (nucleus), and max tokens. All parameters were set uniformly for this study. The temperature was set at a low value of 0.01 to prioritize predictability in the model's output. This setting reduces the randomness and creativity of the responses, favoring deterministic and consistent results ideal for medical diagnostics where accuracy is paramount. The top-P parameter was set at 1, allowing for the broadest selection of words while maintaining focus on relevant content, crucial for generating precise differential diagnoses. Lastly, the max tokens were limited to streamline the output, ensuring that the AI focuses on generating concise, relevant differential-diagnosis lists. Table 1 illustrates the key characteristics of the methods to generate differentials, including adjustable parameters and the prompts. The details of methods to generate differentials, including adjustable parameters and system prompts are shown in Multimedia Appendix 2.

**Table 1.** The key characteristics of the methods to generate differentials include adjustable parameters and the prompts in this study.

| | LLaMA3 | LLaMA2 |
|---|---|---|
| Developer | Meta AI | Meta AI |
| Version | 70B | 70B |
| Release date | April 2024 | July 2023 |
| Access date | May 2024 | May 2024 |
| Prompt | "Tell me the top 10 suspected illnesses for the following case: (copy and paste the case)" | "Tell me the top 10 suspected illnesses for the following case: (copy and paste the case)" |
| Temperature | 0.01 | 0.01 |
| Max tokens | 500 | 500 |
| Top-P | 1 | 1 |

### Evaluation

Two expert physicians, T Shiraishi and T Suzuki, independently evaluated the differentials. We adopted a binary approach to evaluate whether the final diagnosis was included in the differential diagnosis lists. When the lists included the final diagnosis, their rankings were also evaluated. To ensure consistency and objectivity in evaluations, any discrepancies between the initial assessments by T Shiraishi and T Suzuki were resolved through a consensus meeting involving a third expert physician, KT. To enhance the reliability of our evaluation process, we considered implementing a κ statistic to quantify interevaluator agreement. All evaluators were blinded to which AI system generated the differentials to prevent bias. The details of evaluation methods are shown in Multimedia Appendix 3.

### Analysis

In this study, diagnostic performance was defined as the inclusion of the final diagnosis in the differential diagnosis lists.

### Outcome

We defined the primary outcome as the ratio of cases where the final diagnosis was included in the top 10 differential diagnosis lists generated by LLaMA2 or LLaMA3. The denominator was the total number of cases. The numerator was the number of cases in which the final diagnosis was included in the lists. The secondary outcomes were defined as the ratios of whether the final diagnosis was included in the top 5 differential diagnosis lists and as the top diagnosis, generated by LLaMA2 or LLaMA3. We defined the primary outcome and the secondary outcomes as overall diagnostic performance. Additionally, interrater reliability between the physicians' evaluation for the differential diagnosis lists was calculated as the Cohen κ coefficient.

### Exploratory Analysis

The dataset for this analysis comprised cases sourced from a broad spectrum of medical specialties. Each case report was tagged with one to six relevant medical specialties, ensuring a comprehensive representation of the diverse areas in medicine. These specialties were included as part of the standardized metadata attached to each case report, facilitating an organized and targeted analysis. In this study, we included only those specialties that were tagged in at least 10 different case reports.

The exploratory analysis involved quantifying the number of cases correctly diagnosed within each specialty and calculating the ratio of cases for each specialty where the final diagnosis was included in the top 10 differential diagnosis lists generated by LLaMA3 or LLaMA2. The denominator was the total number of cases for each specialty. The numerator was the number of cases in which the final diagnosis was included in the lists. Additionally, we calculated 95% CIs for each ratio to assess the precision of our estimates.

### Statistical Analysis

Data were analyzed using R software (version 4.2.2; The R Foundation for Statistical Computing). Descriptive statistics for categorical or binary variables were summarized as numbers and percentages. We compared categorical data using the chi-square test to determine statistical significance. All tests were 2-sided, and a $P$ value of <.05 was considered to indicate statistical significance. For assessing interrater reliability, the Cohen κ coefficient was computed using the "irr" package in R. The interpretation of Cohen κ was as follows: a value below 0.4 indicates poor agreement; values between 0.41 and 0.75 suggest fair to good agreement; and values from 0.75 to 1 reflect very good to excellent agreement [20].

## Results

### Overview Cases

A total of 392 cases were included in this study. LLaMA3 and LLaMA2 generated differentials for all cases. The representative differentials are shown in Table 2. The dataset of cases, differentials by LLaMA3 and LLaMA2, and the final diagnoses are shown in Multimedia Appendix 4. The cases encompassed a broad range of medical specialties, with the highest number of cases in general internal medicine and surgery, as detailed in Table 3.

**Table 2.** The representative differentials generated by LLaMA3 and LLaMA2.

| Final diagnosis | LLaMA3 | LLaMA2 |
|---|---|---|
| Dementia with Lewy bodies | 1. Lewy body dementia<br>2. Parkinson disease<br>3. Dementia with Lewy bodies<br>4. Alzheimer disease<br>5. Rapid eye movement sleep behavior disorder<br>6. Neuroleptic malignant syndrome<br>7. Serotonin syndrome<br>8. Delirium<br>9. Frontotemporal dementia<br>10. Vascular dementia | 1. Parkinson disease<br>2. Dementia with Lewy bodies<br>3. Alzheimer disease<br>4. Vascular dementia<br>5. Frontotemporal dementia<br>6. Huntington disease<br>7. Creutzfeldt-Jakob disease<br>8. Wernicke-Korsakoff syndrome<br>9. Mild cognitive impairment<br>10. Schizophrenia |
| Infectious myositis | 1. Necrotizing fasciitis<br>2. Sepsis<br>3. Cellulitis<br>4. Myositis<br>5. Pyomyositis<br>6. Abscess<br>7. Osteomyelitis<br>8. Endocarditis<br>9. Bacteremia<br>10. Gas gangrene | 1. Infectious shock<br>2. Sepsis<br>3. Bacterial arthritis<br>4. Muscle abscess<br>5. Necrotizing fasciitis<br>6. Cellulitis<br>7. Pyomyositis<br>8. Osteomyelitis<br>9. Septicemia<br>10. Multiorgan dysfunction syndrome |
| Wandering spleen | 1. Wandering spleen<br>2. Splenoptosis<br>3. Intestinal malrotation<br>4. Pancreatic heterotopy<br>5. Gastric volvulus<br>6. Midgut volvulus<br>7. Internal hernia<br>8. Ovarian torsion<br>9. Uterine fibroid<br>10. Pelvic inflammatory disease | 1. Chronic pancreatitis<br>2. Pancreatic neuroendocrine tumor<br>3. Splenic tumor<br>4. Intestinal obstruction<br>5. Chronic appendicitis<br>6. Pelvic inflammatory disease<br>7. Endometriosis<br>8. Ovarian cyst<br>9. Gastrointestinal stromal tumor<br>10. Mesenteric ischemia |

**Table 3.** Medical specialties in all cases and those correctly identified solely by LLaMA3 and LLaMA2.

| Medical specialty[a] | All cases (N=392), n | Cases correctly identified solely by LLaMA3 (N=312), n (%, 95% CIs) | Cases correctly identified by LLaMA2 (N=195), n (%, 95% CIs); improvement ratios |
|---|---|---|---|
| Surgery | 67 | 50 (74.6, 72.5-77) | 34 (50.7, 49-52.5) |
| General internal medicine | 64 | 55 (85.9, 83.7-88.2) | 35 (54.7, 52.9-56.5) |
| Infectious diseases | 55 | 48 (87.3, 84.8-89.7) | 39 (70.9, 68.7-73.1) |
| Cardiology | 49 | 38 (77.6, 75.1-80) | 20 (40.8, 39.1-42.6) |
| Neurology | 42 | 37 (88.1, 85.3-90.9) | 23 (54.7, 52.5-57) |
| Urology | 40 | 34 (85, 82.1-87.9) | 23 (57.5, 55.2-59.8) |
| Oncology | 32 | 26 (81.3, 78.1-84.4) | 19 (59.4, 56.7-62) |
| Metabolic diseases | 32 | 26 (81.3, 78.1-84.4) | 19 (59.4, 56.7-62) |
| Radiology | 29 | 20 (69, 65.9-72) | 16 (55.2, 52.5-57.9) |
| Critical care medicine | 27 | 22 (81.5, 78.1-84.9) | 9 (33.3, 31.2-35.5) |
| Gastrointestinal | 27 | 21 (77.8, 74.5-81.1) | 10 (37, 34.7-39.3) |
| Hematology | 22 | 17 (77.3, 73.6-80.9) | 10 (45.5, 42.6-48.3) |
| Rheumatology | 19 | 14 (73.7, 69.8-77.5) | 12 (63.2, 59.6-66.7) |
| Nephrology | 18 | 14 (77.8, 73.7-81.9) | 8 (44.4, 41.4-47.5) |
| Respiratory | 18 | 15 (83.3, 79.1-87.6) | 11 (61.1, 57.5-64.7) |
| Obstetrics and gynecology | 17 | 11 (64.7, 60.9-68.5) | 8 (47.1, 43.8-50.3) |
| Endocrinology | 16 | 14 (87.5, 82.9-92.1) | 7 (43.8, 40.5-47) |
| Otolaryngology | 13 | 10 (76.9, 72.2-81.7) | 4 (30.8, 27.8-33.8) |
| Orthopedics | 10 | 6 (60, 55.2-64.8) | 4 (40, 36.1-43.9) |

[a]Each case report was tagged with one to six relevant medical specialties.

## Overall Diagnostic Performance

The final diagnosis was included in the top 10 differentials generated by LLaMA3 in 79.6% (312/392) of the cases, compared to 49.7% (195/392) for LLaMA2, indicating a statistically significant improvement ($P$<.001). Additionally, LLaMA3 showed higher performance in including the final diagnosis in the top 5 differentials, observed in 63% (247/392) of cases, compared to LLaMA2's 38% (149/392, $P$<.001). Moreover, the final diagnosis was accurately identified as the top diagnosis by LLaMA3 in 33.9% (133/392) of cases, significantly higher than the 22.7% (89/392) achieved by LLaMA2 ($P$<.001). The overall diagnostic performance of LLaMA3 and LLaMA2 is shown in Table 4. We observed fair to good agreement between physicians' evaluations for the differential diagnosis lists, with a κ coefficient of 0.69, indicating concordance in 84.2% (660/784) of cases.

**Table 4.** Overall diagnostic performance of LLaMA3 and LLaMA2.

| Diagnostic performance | LLaMA3, n/N (%) | LLaMA2, n/N (%) | $P$ value[a] |
|---|---|---|---|
| The ratio of whether the final diagnosis was included in the top 10 differential diagnosis lists | 312/392 (79.6) | 195/392 (49.7) | <.001 |
| The ratio of whether the final diagnosis was included in the top 5 differential diagnosis lists | 247/392 (63) | 149/392 (38) | <.001 |
| The ratio of whether the final diagnosis was included as the top diagnosis | 133/392 (33.9) | 89/392 (22.7) | <.001 |

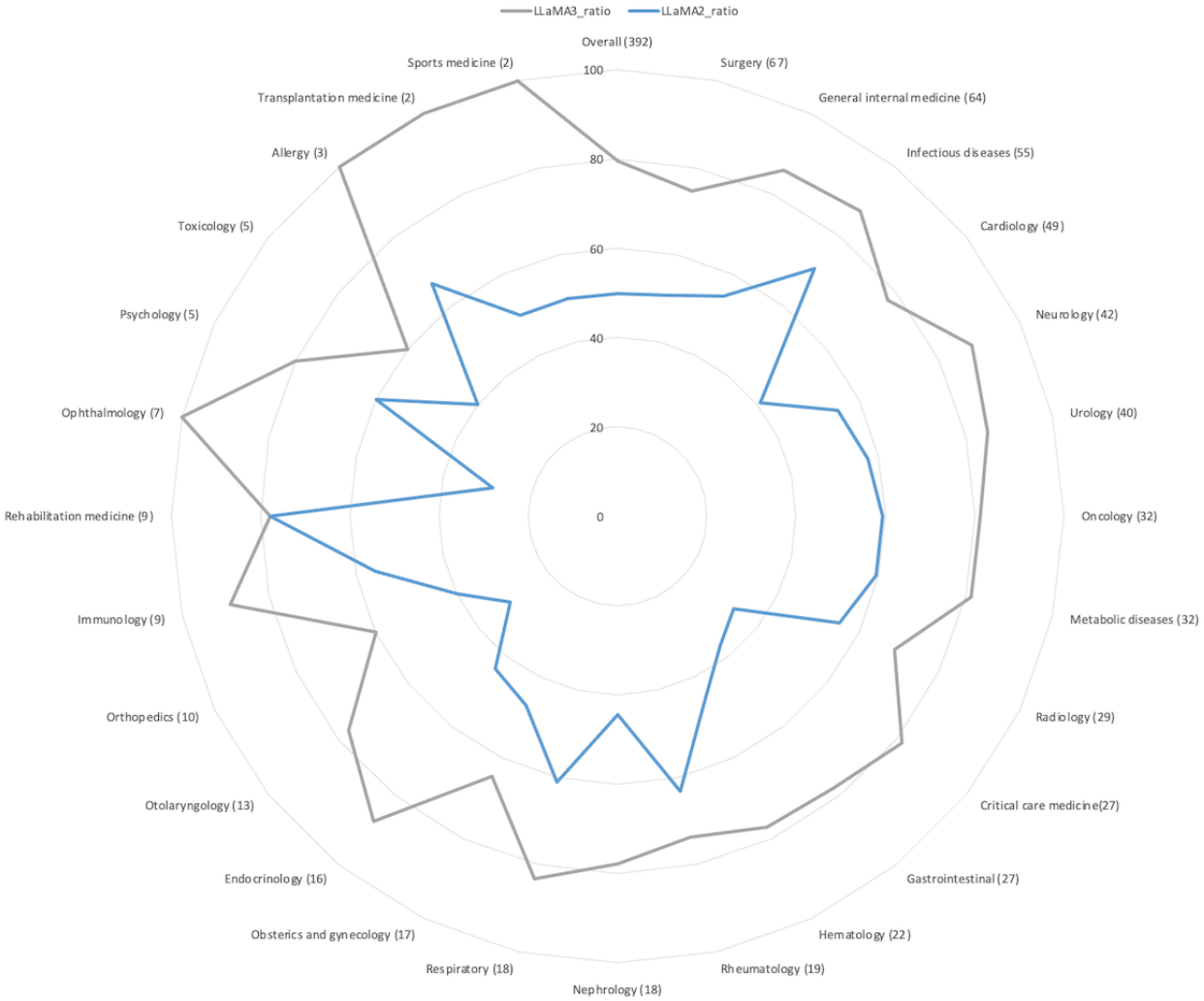
[a]$P$ value from chi-squared test.

## Exploratory Analysis by Medical Specialty

The exploratory analysis across various medical specialties revealed variations in diagnostic performance with LLaMA3 consistently outperforming LLaMA2 in almost all fields. All specialties showed improvements of more than 10% from LLaMA2 to LLaMA3, with nonoverlapping 95% CIs, indicating statistically significant enhancements. Specifically, critical care medicine, gastrointestinal, endocrinology, and otolaryngology exhibited remarkable improvements of more than 40% from LLaMA2. Conversely, infectious diseases, radiology, and obstetrics and gynecology showed the least improvements, with

about a 10% increase from LLaMA2 to LLaMA3. Other specialties exhibited moderate improvements with 20%-30%. Ophthalmology demonstrated the highest accuracy with 71.4% (5/7) of cases correctly identified, followed by otolaryngology at 61.5% (8/31). Lower accuracy was observed in specialties such as rehabilitation medicine at 11.1% (1/9) and rheumatology at 15.8% (3/19). Other specialties such as general internal medicine and surgery showed moderate performance with accuracies of 34.4% (22/64) and 28.4% (19/67), respectively. Table 3 details the breakdown of medical specialties, showing the total number of cases and those correctly identified by LLaMA3 and LLaMA2 in all cases and those correctly identified solely by LLaMA3. Figure 2 presents a radar chart illustrating the ratio of cases for each specialty where the final diagnosis was included in the top 10 differential diagnosis lists generated by both LLaMA3 or LLaMA2.

**Figure 2.** Radar chart illustrating the improvement ratios for the inclusion of the final diagnosis within the top 10 differential diagnosis lists generated by LLaMA2 and LLaMA3, across various medical specialties. Each axis on the radar chart represents a specific medical specialty. The numerical values adjacent to each specialty name reflect the total number of cases analyzed within that specialty, providing context for the observed performance metrics.




## Discussion

### Principal Results

This study demonstrated that the LLaMA3 model significantly outperforms LLaMA2 in overall diagnostic performance, showing almost 1.5-fold improvement. Specifically, the inclusion rate of the final diagnosis in the top 10 differentials rose from 50% to 80%. This substantial enhancement reflects marked advancements within the LLaMA series over a relatively short period.

These enhancements likely come from the implementation of more advanced algorithms and more robust training datasets, highlighting the rapid evolution of generative AI capabilities in medical diagnostics. The significantly higher inclusion rates of the final diagnosis in the top 10, top 5 differentials, and the top diagnosis by LLaMA3 indicate that its model has been finely tuned for greater precision in analyzing complex medical cases. This tuning suggests that LLaMA3 is more adept at incorporating clinical nuances and recognizing a diverse range of symptoms, which is critical for generating accurate differential diagnoses in real-world clinical settings.

### Model Bias and Generalizability

While this study leverages data from a single journal, it is crucial to consider how this might limit the generalizability of the findings. The cases predominantly represent complex or rare medical scenarios, which might not fully represent routine clinical situations found across diverse health care systems [21].

This focus could skew the AI's performance, suggesting that while LLaMA3 shows promise, its effectiveness in general practice remains to be validated in a more varied clinical context.

### Practical Implementation and Challenges

Integrating LLaMA3 into clinical practice presents several challenges that require careful consideration. The foremost is regulatory approval, as generative AI, including the LLaMA series, has not yet been approved for direct clinical applications such as AI-enhanced diagnostics. Regulatory hurdles can significantly delay or impede the practical application of innovative technologies. Furthermore, clinician trust in AI decision-making is vital and requires the AI to be not only effective but also transparent in how decisions are derived. Clinicians must be able to comprehend how decisions are derived to confidently integrate AI recommendations into their workflow.

The computational demands of running sophisticated models such as LLaMA3 also pose a significant challenge. High-performance computing resources, such as Graphics Processing Units or cloud-based solutions, are essential to operate these advanced AI systems effectively, which could limit their deployment in resource-constrained settings.

### Future Research and Development

To facilitate the effective integration of AI such as LLaMA3 into health care workflows, ongoing training with real-world data and continuous feedback from clinical use are indispensable. This iterative process will help ensure that the AI remains accurate and adapts to evolving medical standards. Exploring multimodal AI that incorporates text and image data from electronic health records could enhance diagnostic accuracy. Future studies should focus on integrating these systems with routine health care workflows to assess their practical utility and acceptance among health care providers. Additionally, addressing potential biases in AI decision-making and ensuring adherence to ethical health care standards are crucial for gaining acceptance and trust in clinical environments.

### Results From Exploratory Analysis

The exploratory analysis across different medical specialties provided a view of LLaMA3's performance, which varied across fields. For instance, specialties, including critical care medicine, showed exceptionally high improvements in diagnostic accuracy with LLaMA3. This finding highlights its effectiveness in processing complex clinical courses.

However, the analysis also uncovered areas with modest improvements. For instance, radiology showed small improvements, with about a 10% increase from LLaMA2. This result suggests a need for multimodal AI that can process image data in addition to text data [22]. Multimodal AI enables the simultaneous processing and understanding of multiple forms, including text and image data, which is particularly pertinent for enhancing diagnostic accuracy in radiology.

The variability in these improvements highlights the importance of targeted algorithmic training tailored to the specific demands of each medical specialty. Specialized training datasets that encompass the wide range of scenarios encountered in particular fields could be crucial in enhancing the generative AI's learning curve and improving its utility in clinical practice. The performance of LLaMA3 varies across medical specialties, with notably high improvement ratios in ophthalmology and otolaryngology, likely due to the distinct and well-defined symptoms associated with conditions treated within these fields. Conversely, specialties such as rehabilitation medicine and rheumatology showed lower improvement ratios, attributed to the complexity of the clinical course and immune responses, posing challenges for the current model's diagnostic algorithms. A significant factor contributing to the variation in performance is the relatively small number of cases available for some specialties.

### Strengths

A major strength of this study is the controlled comparison of diagnostic performances using identical cases and standardized parameters, providing a clear assessment of improvements from LLaMA2 to LLaMA3. Additionally, the longitudinal assessment of the LLaMA series offers valuable insights into the developmental course of AI models in medical diagnostics. This is particularly notable when contrasted with findings from other AI systems where no improvement was noted over time [23].

### Limitations

#### *Overview*

There were several limitations concerning study design and generative AI.

#### *Limitations for Study Design*

First, case reports may not fully reflect real-world clinical cases. This limitation arises because case reports often focus on new or rare diseases, which might not be commonly encountered in typical clinical settings [21]. Second, relying solely on a single case report journal may introduce selection bias. Third, there was no well-established standard to evaluate the diagnostic performance of clinical decision support systems, including the number of differentials and the evaluation methods. For example, a study adopted 5 differentials while another adopted 40 differentials [24,25]. Regarding evaluation methods, some studies used scale-based assessments, while others used binary methods. Qualitative evaluations of the differential diagnosis lists should also be explored in future studies to assess their overall clinical relevance beyond whether the correct diagnosis was included. These variations in evaluation methods were partly due to the complexity of the diagnostic process in real clinical situations [26]. Fourth, we excluded specialties tagged in fewer than 10 different case reports. Therefore, there was a possibility to overlook minor specialties where LLaMA3 did not outperform LLaMA2. Fifth, the variability in sample sizes across specialties in our exploratory analysis might affect the robustness of the conclusions drawn. Additionally, the sensitivity of AI models such as the LLaMA series to variations in input prompts—prompt engineering—is a critical area. There is a potential that even minor prompt changes presented to the AI can significantly influence its diagnostic suggestions, emphasizing the need for standardized prompt protocols to ensure consistent AI performance.

### *Limitations for Generative AI*

Generative AI, including the LLaMA series, has not been approved for medical applications such as AI-enhanced diagnostics. Additionally, the optimal prompts and adjustable parameters for medical diagnostics remain unknown. For example, another study used different settings with a temperature of 0.6, top-P of 0.9, and max tokens of 2048 [16], in contrast to our study which used a temperature of 0.01, top-P of 1, and max tokens of 500. Similarly, another study used multiple prompting scenarios, such as chain of thought, few shots, and retrieval augmentations [27], compared to our study with a simple prompt. This difference in prompting complexity could impact the generative AI's performance. Furthermore, we did not recruit all available generative AI, including the ChatGPT series, Gemini, and Claude 3. Moreover, a critical limitation identified in our study involves the potential for data leakage, where LLaMA3 and LLaMA2 might have been previously exposed to the case reports used in our analysis, thereby influencing their performance artificially. The inherent risk of data leakage cannot be entirely ruled out due to the models' continuous learning capabilities and the complex nature of their training environments. To mitigate such risks in future studies, we plan to implement rigorous partitioning of data to ensure that no overlap occurs between training and testing datasets. Regarding transparency, although the LLaMA series is often referred to as open-source LLMs, there is an ongoing debate about the openness of generative AIs [28,29]. Finally, the rapid pace of development in generative AI systems suggested that our findings may quickly become outdated as next-generation LLMs emerge.

These limitations could affect generalizability.

## Comparison With Prior Work

### *Comparison With LLaMA2*

Following the limitations outlined, our comparative analysis with prior iterations of LLaMA2 highlights the dynamic nature of AI development and its implications on diagnostic accuracy. In our study, the inclusion of the final diagnosis in the top 10 differentials for 49.7% (195/392) of cases represents a decrease from the 54.6% (214/392) observed in our prior study [18]. This variation in performance, a 1%-5% difference, is directly attributable to the adjustments in operational parameters such as temperature, max tokens, and top-P. These findings highlight how seemingly minor tweaks in AI configurations can lead to significant changes in outcome, emphasizing the necessity for continuous optimization based on evolving clinical needs.

Our results not only reflect the critical impact of parameter adjustments on the efficacy and reliability of AI diagnostic outputs but also the importance of tailoring these settings to specific diagnostic tasks within clinical environments. The ongoing research and development efforts are vital as they contribute to refining these parameters to enhance the performance of AI systems in real-world settings. Table 5 details the diagnostic performance and key characteristics of LLaMA2 compared to the previous study, illustrating these points and showing the progression within the LLaMA series.

**Table 5.** Diagnostic performance and key characteristics of LLaMA2 compared to a previous study.

| | LLaMA2 in this study | LLaMA2 in the previous study |
|---|---|---|
| The ratio of whether the final diagnosis was included in the top 10 differential diagnosis lists, n/N (%) | 195/392 (49.7) | 214/392 (54.6) |
| The ratio of whether the final diagnosis was included in the top 5 differential diagnosis lists, n/N (%) | 149/392 (38) | 177/392 (45.2) |
| The ratio of whether the final diagnosis was included as the top diagnosis, n/N (%) | 89/392 (22.7) | 90/392 (23) |
| Developer | Meta AI | Meta AI |
| Version | 70B | 70B |
| Release date | July 2023 | July 2023 |
| Access date | May 2024 | August 2023 |
| Prompt | "Tell me the top 10 suspected illnesses for the following case: (copy and paste the case)" | "Tell me the top 10 suspected illnesses for the following case: (copy and paste the case)" |
| Temperature | 0.01 | 2.49 |
| Max tokens | 500 | 2048 |
| Top-P | 1 | 0.5 |

### *Comparison With Other Generative AI*

From another study involving ChatGPT-3.5, ChatGPT-4, and LLaMA2, inferior performances of LLaMA2 compared to ChatGPT-3.5 and ChatGPT-4 were observed [16]. From the current findings, there is a possibility that results may change due to longitudinal improvements from LLaMA2 to LLaMA3.

Our comparative analysis extends beyond LLaMA2 and LLaMA3 to include contemporary models such as ChatGPT-4 and Gemini, providing a broader perspective on generative AI capabilities. While LLaMA3 has shown notable improvements and closely matches the performance of ChatGPT-4 with a diagnostic accuracy of 86.7% (340/392) in the top 10 differentials [18], it is essential to consider the development timelines and the operational models of these AI systems. Unlike LLaMA3, ChatGPT-4(o) and Gemini Advanced are fee-based models that might have different optimization and deployment strategies, potentially affecting their performance in clinical settings. Moreover, the introduction of newer models such as ChatGPT-4o and OpenAI o1 represents continuous advancements within the generative AI landscape, highlighting the dynamic nature of AI development.

### *Comparison With Other Clinical Decision Support Systems*

Expanding on our comparative analysis, we also evaluate LLaMA3 in the context of established clinical decision support systems such as Isabel Pro (developed by Isabel Healthcare). While Isabel Pro has demonstrated a diagnostic retrieval accuracy of 65% for its top 10 differentials, increasing to 87% for the top 40 [25], these figures provide a benchmark for evaluating LLaMA3's capabilities. Our study's performance metrics are closely aligned with these established systems, suggesting that LLaMA3 could offer comparable benefits in clinical decision-making. It is crucial to understand the methodologies and metrics used across different systems to ensure a fair and meaningful comparison.

## Conclusions

The results demonstrate that the LLaMA3 model significantly outperforms LLaMA2 per diagnostic performance, with a higher percentage of case reports having the final diagnosis listed within the top 10, top 5, and as the top diagnosis. Overall diagnostic performance improved almost 1.5 times from LLaMA2 to LLaMA3. These findings support the rapid development and continuous refinement of generative AI systems to enhance diagnostic processes in medicine. However, these findings should be carefully interpreted for clinical application, as generative AI, including the LLaMA series, has not been approved for medical applications such as AI-enhanced diagnostics.

## Acknowledgments

This research was funded by JSPS (Japan Society for the Promotion of Science) KAKENHI (grant 22K10421). This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

## Authors' Contributions

TH, YH, KT, T Shiraishi, T Suzuki, and T Shimizu contributed to this study's concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. YH, KT, T Shiraishi, T Suzuki, and T Shimizu contributed to the critical revision of the manuscript for relevant intellectual content. All the authors have read and approved the final version of the manuscript.

## Conflicts of Interest

The author reports no conflicts of interest in this work, except for T Shiraishi, who Ubie, Inc employs.

## Multimedia Appendix 1

The details of preparing case reports.
[DOCX File , 21 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

The details of methods to generate differentials, including adjustable parameters and system prompt.
[DOCX File , 21 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

The details of evaluation methods.
[DOCX File , 20 KB-Multimedia Appendix 3]

## Multimedia Appendix 4

The dataset of cases, differentials, and the final diagnoses, used in our study.

[XLSX File (Microsoft Excel File), 149 KB-Multimedia Appendix 4]

## References


1. Turing AM. Computing machinery and intelligence. Mind. 1950;59:433-460. [doi: 10.1093/oso/9780198250791.003.0017]
2. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: a survey. Heliyon. 2018;4(11):e00938. [FREE Full text] [doi: 10.1016/j.heliyon.2018.e00938] [Medline: 30519653]
3. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y. A survey of large language models. arXiv:2303.18223. Preprint published online on March 31, 2023
4. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568. [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
5. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
6. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJPC. Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access. 2024;12:31078-31106. [doi: 10.1109/access.2024.3367715]
7. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform. 2024;12:e53787. [FREE Full text] [doi: 10.2196/53787] [Medline: 38728687]
8. Balas M, Micieli JA. Visual snow syndrome: use of text-to-image artificial intelligence models to improve the patient perspective. Can J Neurol Sci. 2023;50(6):946-947. [doi: 10.1017/cjn.2022.317] [Medline: 36352764]
9. Restrepo D, Rodman A, Abdulnour R. Conversations on reasoning: large language models in diagnosis. J Hosp Med. 2024;19(8):731-735. [doi: 10.1002/jhm.13378] [Medline: 38678438]
10. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. 2024;184(5):581-583. [doi: 10.1001/jamainternmed.2024.0295] [Medline: 38557971]
11. OpenAI. GPT-4 technical report. OpenAI. 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2024-10-22]
12. Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, et al. Peer review of GPT-4 technical report and systems card. PLOS Digit Health. 2024;3(1):e0000417. [FREE Full text] [doi: 10.1371/journal.pdig.0000417] [Medline: 38236824]
13. Gemini Team Google. Gemini: a family of highly capable multimodal models. arXiv:2312.11805. Preprint published online on December 19, 2023
14. Meta. Build the future of AI with Meta Llama 3 2024. URL: https://llama.meta.com/llama3/ [accessed 2024-10-22]
15. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80. [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
16. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and llama 2 for clinical decision support tasks. Nat Commun. 2024;15(1):2050. [FREE Full text] [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
17. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA. 2024;331(15):1320-1321. [doi: 10.1001/jama.2023.27861] [Medline: 38497956]
18. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. Digit Health. 2024;10:20552076241265215. [FREE Full text] [doi: 10.1177/20552076241265215] [Medline: 39229463]
19. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med. 2008;23 Suppl 1(Suppl 1):37-40. [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
20. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. New York. John Wiley & Sons; 2003.
21. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. J Clin Epidemiol. 2017;89:218-235. [doi: 10.1016/j.jclinepi.2017.04.026] [Medline: 28529185]
22. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
23. Harada Y, Sakamoto T, Sugimoto S, Shimizu T. Longitudinal changes in diagnostic accuracy of a differential diagnosis list developed by an AI-based symptom checker: retrospective observational study. JMIR Form Res. 2024;8:e53985. [FREE Full text] [doi: 10.2196/53985] [Medline: 38758588]
24. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med. 2024;83(1):83-86. [doi: 10.1016/j.annemergmed.2023.08.003] [Medline: 37690022]

25. Bridges JM. Computerized diagnostic decision support systems—a comparative performance study of Isabel Pro vs. ChatGPT4. Diagnosis (Berl). 2024;11(3):250-258. [FREE Full text] [doi: 10.1515/dx-2024-0033] [Medline: 38709491]
26. Merkebu J, Battistone M, McMains K, McOwen K, Witkop C, Konopasky A, et al. Situativity: a family of social cognitive theories for understanding clinical reasoning and diagnostic error. Diagnosis (Berl). 2020;7(3):169-176. [FREE Full text] [doi: 10.1515/dx-2019-0100] [Medline: 32924378]
27. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y). 2024;5(3):100943. [FREE Full text] [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
28. Liesenfeld A, Dingemanse M. Rethinking open source generative AI: open-washing and the EU AI Act. 2024. Presented at: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; 2024 June 05:1774-1787; Rio de Janeiro, Brazil. [doi: 10.1145/3630106.3659005]
29. WHO. Ethics and governance of artificial intelligence for health. USA. WHO guidance; 2021. URL: https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1 [accessed 2024-10-17]

## Abbreviations

**AI:** artificial intelligence
**LLM:** large language model
**NLP:** natural language processing

*Edited by A Mavragani; submitted 28.07.24; peer-reviewed by S Mao, M Popovic; comments to author 19.09.24; revised version received 23.09.24; accepted 01.10.24; published 19.11.24*

*Please cite as:*
*Hirosawa T, Harada Y, Tokumasu K, Shiraishi T, Suzuki T, Shimizu T*
*Comparative Analysis of Diagnostic Performance: Differential Diagnosis Lists by LLaMA3 Versus LLaMA2 for Case Reports*
*JMIR Form Res 2024;8:e64844*
*URL: https://formative.jmir.org/2024/1/e64844*
*doi: 10.2196/64844*
*PMID:*