JMIR Publications
The Leading eHealth Publisher

Comparative Analysis of Diagnostic Performance: Differential Diagnosis Lists by LLaMA3 versus LLaMA2 for case reports

Takanobu Hirosawa, Yukinori Harada, Kazuki Tokumasu, Tatsuya Shiraishi, Tomoharu Suzuki, Taro Shimizu

Submitted to: JMIR Formative Research
on: July 28, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

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] MD, PhD, MSc, MPH, MBA, FACP

[1]Dokkyo Medical University Department of Diagnostic and Generalist Medicine Shimotsuga JP
[2]Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences Department of General Medicine Okayama JP
[3]Higashinihonbashinaika clinic Tokyo JP
[4]Ubie, inc Tokyo JP
[5]Urasoe General Hospital Department of Hospital Medicine Okinawa JP

Corresponding Author:
Takanobu Hirosawa MD, PhD
Dokkyo Medical University
Department of Diagnostic and Generalist Medicine
880 Kitakobayashi, Mibu-cho
Shimotsuga
JP

Preprint JMIR Publications

Abstract

Background: Generative artificial intelligence (AI), particularly in the form of large language models (LLMs), has rapidly developed. The LLM by Meta AI (LLaMA) series are popular and recently updated from LLaMA2 to LLaMA3. However, impacts of the update in 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 non-diagnostic 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 63.0% (247/392) of cases, compared to LLaMA2's 38.0% (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 in terms of 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 dDiagnostic 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. Clinical Trial: Not applicable

(JMIR Preprints 28/07/2024:64844)
DOI: https://doi.org/10.2196/preprints.64844

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint JMIR Publications

Original Paper

Comparative Analysis of Diagnostic Performance: Differential Diagnosis Lists by LLaMA3 versus LLaMA2 for case reports

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

Affiliations:
1 Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, 321-0293, Japan.
2 Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan.
3 Higashinihonbashinaika clinic / Ubie, inc
4 Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan.

Correspondence: Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University, 880 Kitakobayashi, Mibu-cho, Shimotsuga, Tochigi, Japan 321-0293
Tel + 81-282-87-2498
Fax + 81-282-87-2502
Email: hirosawa@dokkyomed.ac.jp

Preprint JMIR Publications

Abstract

Background: Generative artificial intelligence (AI), particularly in the form of large language models (LLMs), has rapidly developed. The LLM by Meta AI (LLaMA) series are popular and recently updated from LLaMA2 to LLaMA3. However, impacts of the update in 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 non-diagnostic 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 63.0% (247/392) of cases, compared to LLaMA2's 38.0% (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 in terms of 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 dDiagnostic 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.

Trial Registration: Not applicable

Keywords: Artificial Intelligence; Clinical Decision Support System; Generative Artificial Intelligence; Large Language Models; Natural Language Processing

JMIR Preprints

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, 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]. Natural language processing (NLP) enables computers to understand and process human language, facilitating tasks like 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, like the GPT series developed by OpenAI, Google's Gemini, and LLM by Meta AI (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 recently to GPT-4o [11, 12].

Similarly, other systems like 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, like 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 like 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 flow chart, including preparing case reports and generating differentials, is shown in Figure 1.

Ethical Approval

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 non-diagnostic 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. The final diagnoses were typically written by the authors. The main investigator, Takanobu Hirosawa, conducted this process, which was validated by another co-investigator, Yukinori Harada. Details of preparing case reports are shown in Multimedia Appendix 1.

Differentials Generated by Artificial Intelligences

We employed 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 top 10 differential diagnosis lists.

Both LLaMA3 and LLaMA2 allowed for several adjustable settings to control the output, including temperature, top P, and max tokens. All parameters were set uniformly for this study. Temperature settings, adjustable from 0 to 5, control the randomness and creativity of the model's output, with lower settings yielding more predictable results and higher settings increasing variability. Top P, ranging from 0 to 1, adjusts the diversity of the model's predictions by controlling the randomness; a higher value allows a broader selection of the words. Max tokens define the maximum number of word segments, or tokens, that the model can generate in a single output. 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 prompt are shown in Multimedia Appendix 2.

Table 1. The key characteristics of the methods to generate differentials, including 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

LLaMA: large language model by Meta AI

Evaluation

Two expert physicians, Tatsuya Shiraishi and Tomoharu 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. Any discrepancies were resolved by another expert physician, Kazuki Tokumasu. All evaluators were blinded to which AI system generated the differentials. The details of evaluation methods are shown in Multimedia Appendix 3.

Analysis

In this study, diagnostic performance was defined as the differential diagnosis lists included the final diagnosis.

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.

Exploratory Analysis

We conducted an exploratory analysis on the medical specialties associated with cases that were correctly identified by LLaMA3 but not by and LLaMA2. This analysis aimed to find specific areas where LLaMA3 shows enhanced diagnostic capabilities and to understand how advancements in generative AI technology might be contributing differently across various medical fields.

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 calculateding 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% confidence intervals

(CIs) for each ratio to assess the precision of our estimates.the improvement ratios to assess LLaMA3's performance enhancement.

Statistical Analysis

Data were analyzed using R software, version 4.2.2, provided by The R Foundation for Statistical Computing, located in Vienna, Austria. 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 two-sided, and a *P* value <.05 was considered to indicate statistical significance.

Results

Overview Cases

Total 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 4.

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

Final diagnosis	LLaMA3	LLaMA2
Dementia with Lewy bodies		
	1. Lewy Body Dementia 2. Parkinson's Disease 3. Dementia with Lewy Bodies 4. Alzheimer's Disease 5. REM Sleep Behavior Disorder 6. Neuroleptic Malignant Syndrome 7. Serotonin Syndrome 8. Delirium 9. Frontotemporal Dementia 10. Vascular Dementia	1. Parkinson's disease 2. Dementia with Lewy bodies 3. Alzheimer's disease 4. Vascular dementia 5. Frontotemporal dementia 6. Huntington's disease 7. Creutzfeldt-Jakob disease 8. Wernicke-Korsakoff syndrome 9. Mild cognitive impairment 10. Schizophrenia
Infectious myositis		
	1. Necrotizing fasciitis 2. Sepsis 3. Cellulitis 4. Myositis 5. Pyomyositis 6. Abscess 7. Osteomyelitis	1. Infectious shock 2. Sepsis 3. Bacterial arthritis 4. Muscle abscess 5. Necrotizing fasciitis 6. Cellulitis 7. Pyomyositis

Preprint JMIR Publications

	8. Endocarditis 9. Bacteremia 10. Gas gangrene	8. Osteomyelitis 9. Septicemia 10. Multiorgan dysfunction syndrome
Wandering spleen		
	1. Wandering spleen 2. Splenoptosis 3. Intestinal malrotation 4. Pancreatic heterotopy 5. Gastric volvulus 6. Midgut volvulus 7. Internal hernia 8. Ovarian torsion 9. Uterine fibroid 10. Pelvic inflammatory disease	1. Chronic pancreatitis 2. Pancreatic neuroendocrine tumor 3. Splenic tumor 4. Intestinal obstruction 5. Chronic appendicitis 6. Pelvic inflammatory disease 7. Endometriosis 8. Ovarian cyst 9. Gastrointestinal stromal tumor 10. Mesenteric ischemia

LLaMA: large language model by Meta AI

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 63.0% (247/392) of cases, compared to LLaMA2's 38.0% (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). OverallThe diagnostic performance of LLaMA3 and LLaMA2 is shown in Table 3.

Table 3. Overall dDiagnostic performance of LLaMA3 and LLaMA2.

Diagnostic performance	LLaMA3	LLaMA2	*P* value[a]
The ratio of whether the final diagnosis was included in the top 10 differential diagnosis lists, n (%)			
	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, n (%)			
	247/392 (63.0%)	149/392 (38.0%)	<.001
The ratio of whether the final diagnosis was included as top diagnosis, n (%)			
	133/392 (33.9%)	89/392 (22.7%)	<.001

[a] *P* value from chi-squared test.
LLaMA: large language model by Meta AI

Exploratory Analysis by Medical Specialty

Among these cases, 131 were correctly identified only by LLaMA3. The exploratory analysis across various medical specialties revealed varying levelsvariations inof improvement ratiosdiagnostic performance with LLaMA3 consistently outperforming LLaMA2 in almost all fields. All specialties showed improvements of more than 10% from LLaMA2 to LLaMA3, with non-overlapping 95% CIs, indicating statistically significant enhancements. Specifically, Critical Care Medicine, Gastrointestinal, Endocrinology, and Otolaryngology exhibited remarkable improvements of more than 40% form LLaMA2. Conversely, Infectious Diseases, Radiology, and Obstetrics and Gynecology showed the least improvements, with about 10% increase from LLaMA2 to LLaMA2. 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 like general internal medicine and surgery showed moderate performance with accuracies of 34.4% (22/64) and 28.4% (19/67), respectively. Table 4 provides a detailsed 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.

Table 4. Medical specialties in all cases and those correctly identified solely by LLaMA3 and LLaMA2

Medical Specialty[a]	All cases	Cases correctly identified solely by LLaMA3, n (% [95%CIs])	Cases correctly identified by LLaMA2, n (% [95%CIs])Improvement ratios (%)
	N=392	N=312131	N=195
Surgery			
	67	50 (74.6% [72.5-77.0])19	34 (50.7% [49.0-52.5])28.4%

General Internal Medicine			
	64	55 (85.9% [83.7-88.2])22	35 (54.7% [52.9-56.5])4.4%
Infectious Ddiseases			
	55	48 (87.3% [84.8-89.7])11	20.0%39 (70.9% [68.7-73.1])
Cardiology			
	49	38 (77.6% [75.1-80.0])19	20 (40.8% [39.1-42.6])38.8%
Neurology			
	42	37 (88.1% [85.3-90.9])14	23 (54.7% [52.5-57.0])33.3%
Urology			
	40	34 (85.0% [82.1-87.9])12	23 (57.5% [55.2-59.8])30.0%
Oncology			
	32	26 (81.3% [78.1-84.4])9	28.1%19 (59.4% [56.7-62.0])
Metabolic Ddiseases			
	32	26 (81.3% [78.1-84.4])9	19 (59.4% [56.7-62.0])28.1%
Radiology			
	29	20 (69.0% [65.9-72.0])7	16 (55.2% [52.5-57.9])24.1%
Critical Care Medicine Gastrointestinal			
	27	22 (81.5% [78.1-84.9])11	9 (33.3% [31.2-35.5])40.7%
Gastrointestinal Critical Care Medicine			
	27	21 (77.8% [74.5-81.1])13	10 (37.0% [34.7-39.3])48.1%
Hematology			
	22	17 (77.3% [73.6-80.9])9	40.9%10 (45.5% [42.6-48.3])
Rheumatology			
	19	14 (73.7% [69.8-77.5])3	15.8%12 (63.2% [59.6-66.7])
Nephrology Respiratory			
	18	14 (77.8% [73.7-81.9])4	22.2%8 (44.4% [41.4-47.5])
Respiratory Nephrology			
	18	15 (83.3% [79.1-87.6])8	44.4%11 (61.1% [57.5-64.7])
Obstetrics and Gynecology			
	17	11 (64.7% [60.9-68.5])4	23.5%8 (47.1% [43.8-50.3])
Endocrinology			

	16	14 (87.5% [82.9-92.1])7	7 (43.8% [40.5-47.0])43.8%
Otolaryngology			
	13	10 (76.9% [72.2-81.7])8	61.5%4 (30.8% [27.8-33.8])
Orthopedics			
	10	6 (60.0% [55.2-64.8])3	4 (40.0% [36.1-43.9])30.0%
Immunology			
	9	3	33.3%
Rehabilitation Medicine			
	9	1	11.1%
Ophthalmology			
	7	5	71.4%
Psychology			
	5	1	20.0%
Toxicology			
	5	1	20.0%
Allergy			
	3	1	33.3%
Transplantation Medicine			
	2	1	50.0%
Sports Medicine			
	2	1	50.0%

[a]Each case report was tagged with one to six relevant medical specialties
LLaMA: large language model by Meta AI
95%CIs: 95% confidence intervals

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 and, top 5 differentials, as well, and as 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 diverse symptomatology, which are critical for generating accurate differential diagnoses in real-world clinical settings.

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 the 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 [20]. 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 symptomatology of conditions treated within these fields. Conversely, specialties like rehabilitation medicine and rheumatology showed lower improvement ratios, attributed to the complexity of 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 [21].

Limitations

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 [22]. 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 for 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 [23, 24]. Regarding evaluation methods, some studies used scale-based assessments, while others employed binary methods. These variations in evaluation methods were partly due to the complexity of diagnostic process in real clinical situations [25]. 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.

Limitations for Generative Artificial Intelligence

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 employed 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 utilized multiple prompting scenarios, such as chain of thought, few shots, and retrieval augmentation [26], 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, there is a risk that LLaMA3 and LLaMA2 may have already learned from the case reports, potentially biasing the results. Regarding transparency, although the LLaMA series is often referred to as open-source LLMs, there is ongoing debate about the openness of generative AIs [27, 28]. 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 current 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 like 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 previous study.

	LLaMA2 in current study	LLaMA2 in previous study
The ratio of		

whether the final diagnosis was included in the top 10 differential diagnosis lists, 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 (%)		
	149/392 (38.0%)	177/392 (45.2%)
The ratio of whether the final diagnosis was included as top diagnosis, n (%)		
	89/392 (22.7%)	90 (23.0)
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

LLaMA: large language model by Meta AI

Comparison with Other Generative Artificial Intelligence

From another study involving ChatGPT-3.5, ChatGPT-4, and LLaMA2, inferior performances of LLaMA2 compared to ChatGPT-3.5 and ChatGPT-4 was observed (16). From the current findings,

there is a possibility that results may change due to longitudinal improvements from LLaMA2 to LLaMA3.

Compared to ChatGPT-4 [18], our previous study revealed that ChatGPT-4 generated the top 10 differentials with diagnostic performance of 86.7% (340/392). Although LLaMA3 has significantly improved from LLaMA2, it has nearly reached the diagnostic performance of ChatGPT-4, which is noteworthy considering that LLaMA3 is a non-fee model while ChatGPT-4 is a fee-based model. This outcome demonstrates that even non-fee models like the LLaMA series can achieve diagnostic accuracy close to that of fee-based models such as ChaGPT-4. It is important to note that the ChatGPT series was released earlier than the LLaMA series, allowing it more time to incorporate user feedback and undergo iterative improvements. This extended period for enhancements and user engagement is partly why the ChatGPT series has historically had a performance edge, but the rapid catch-up by LLaMA3 highlights significant advancements in the LLaMA series and its potential for further evolution.

Comparison with Other Clinical Decision Support Systems

Compared to Isabel Pro, a successful clinical decision support system developed by Isabel Healthcare, Ltd., a previous study has shown that Isabel Pro's top 10 diagnostic retrieval accuracy was 65%, which increased to 87% for the top 40 differentials [24]. The variations were due to differences in case materials, evaluation methods, and the number of differentials included.

Conclusions

The results demonstrate that the LLaMA3 model significantly outperforms LLaMA2 in terms of 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 dDiagnostic 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.

Acknowledgements

Contributors: TH, YH, KT, TS (Tatsuya Shiraishi), TS (Tomoharu Suzuki), and TS (Taro Shimizu) contributed to the study concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. YH, KT, TS (Tatsuya Shiraishi), TS (Tomoharu Suzuki), and TS (Taro 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.
This research was funded by JSPS KAKENHI (grant number 22K10421). This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial Intelligence
CI: confidence interval
LLaMA: large language model by Meta AI
LLM: large language model
NLP: Natural language processing

Multimedia Appendix 1

The details of preparing case reports.

Multimedia Appendix 2

The details of methods to generate differentials, including adjustable parameters and system prompt.

Multimedia Appendix 3

The details of evaluation methods.

Multimedia Appendix 4

The dataset of cases, differentials, and the final diagnoses, utilized in the current study.

References

1. Turing AM. Computing Machinery and Intelligence. Mind. 1950;59:433-60.
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.
3. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint arXiv:230318223. 2023.
4. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568.
5. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Medical education. 2024.
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;12:31078-106.
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.

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):946-7.
9. Restrepo D, Rodman A, Abdulnour RE. Conversations on reasoning: Large language models in diagnosis. J Hosp Med. 2024.
10. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R-E, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Internal Medicine. 2024;184(5):581-3.
11. OpenAI. GPT-4 Technical Report2023 March 01, 2023:[arXiv:2303.08774 p.].
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 Digital Health. 2024;3(1):e0000417.
13. Team G, Anil R, Borgeaud S, Wu Y, Alayrac J-B, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
14. AI M. Build the future of AI with Meta Llama 3 2024 [Available from: https://llama.meta.com/llama3/.
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.
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.
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):1320-1.
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. DIGITAL HEALTH. 2024;10:20552076241265215.
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.
20. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nature Medicine. 2022;28(9):1773-84.
21. 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.
22. 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. Journal of Clinical Epidemiology. 2017;89:218-35.
23. 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):83-6.
24. Bridges JM. Computerized diagnostic decision support systems – a comparative performance study of Isabel Pro vs. ChatGPT4. Diagnosis. 2024.
25. 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):169-76.
26. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5(3).
27. Liesenfeld A, Dingemanse M. Rethinking open source generative AI: open washing and the EU AI Act. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; Rio de Janeiro, Brazil: Association for Computing Machinery; 2024. p. 1774–87.

28. Organization WH. Ethics and governance of artificial intelligence for health: WHO guidance. 2021.

Preprint JMIR Publications

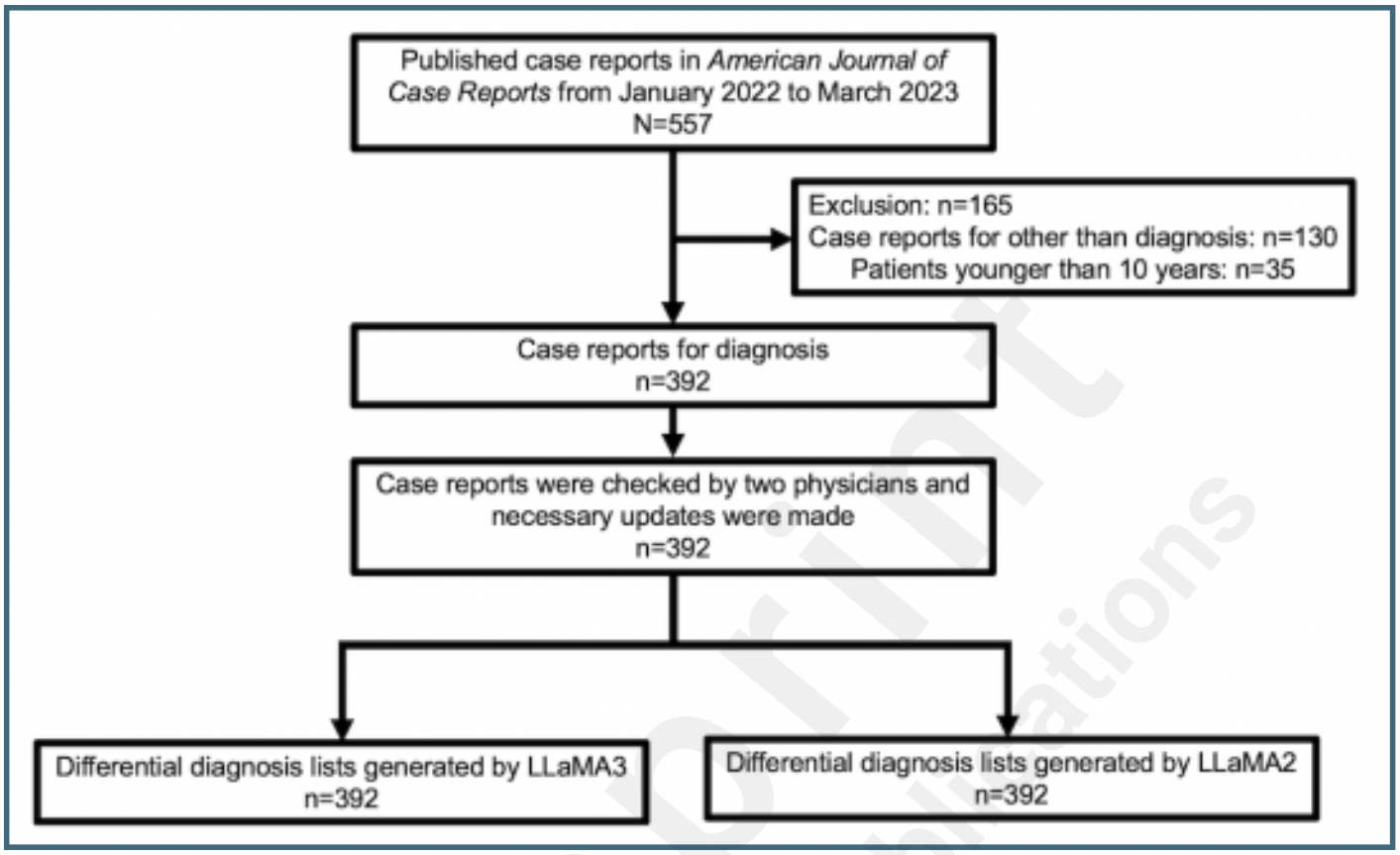
Supplementary Files

Preprint JMIR Publications

Figures

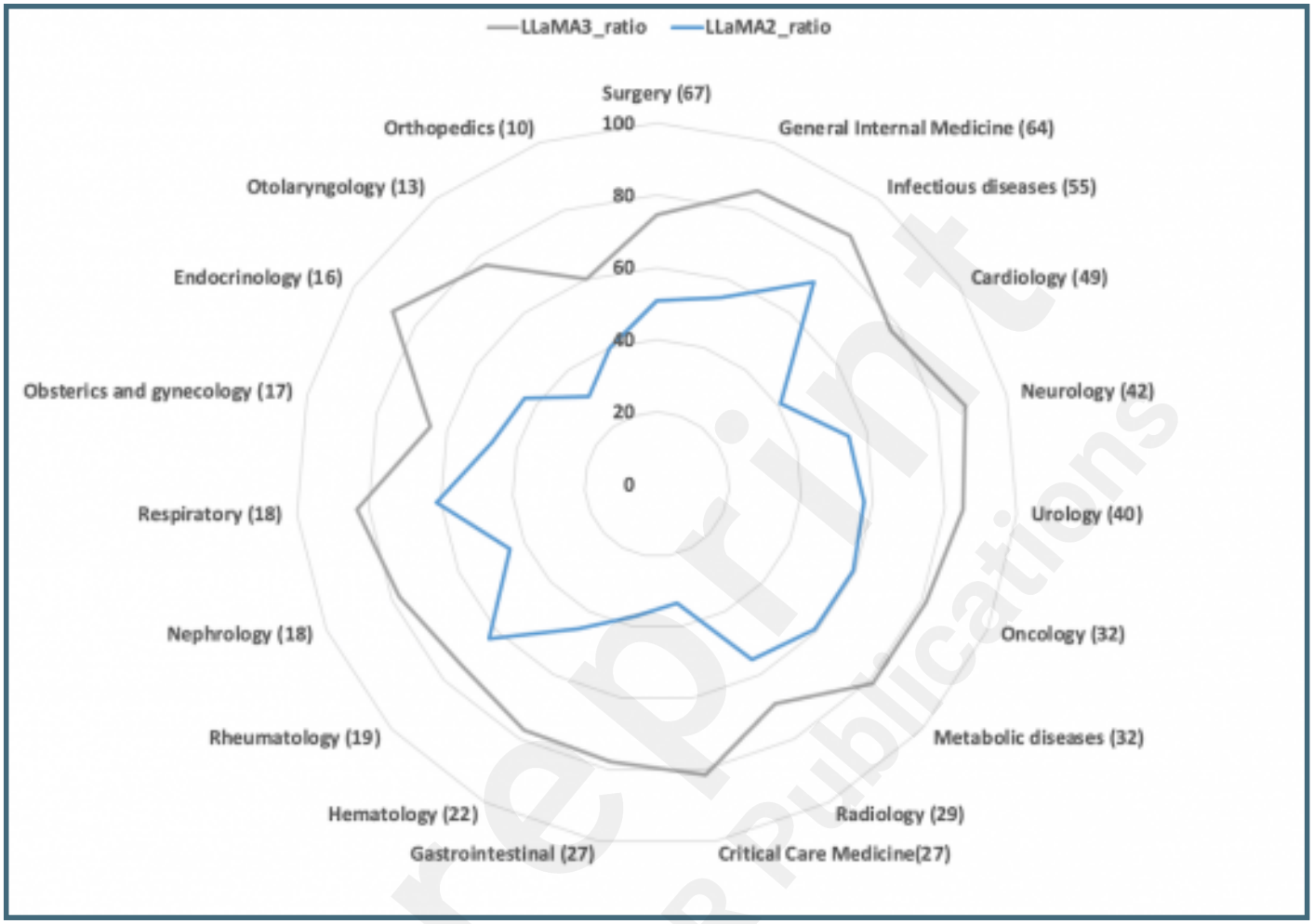
Preprint JMIR Publications

The flow chart, including preparing case reports and generating differentials.



Preprint JMIR Publications

Radar chart for 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 numerical values next to each specialty indicate the number of cases analyzed for the specialty.



Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

The details of preparing case reports.
URL: http://asset.jmir.pub/assets/91d498a52d898fa8abe0973b96621de3.docx

The details of methods to generate differentials, including adjustable parameters and system prompt.
URL: http://asset.jmir.pub/assets/917f6835c4c9ab8657276b1f8a909645.docx

The details of evaluation methods.
URL: http://asset.jmir.pub/assets/060a188e669c14b31a6635405e098608.docx

The dataset of cases, differentials, and the final diagnoses, utilized in the current study.
URL: http://asset.jmir.pub/assets/13a3cd053eccbda8292fba68ade8f990.xlsx

Preprint JMIR Publications