Is it time for the neurologist to use Large Language Models in everyday practice?

Natale Vincenzo Maiorana, Sara Marceglia, Mauro Treddenti, Mattia Tosi, Matteo Guidetti, Maria Francesca Creta, Tommaso Bocci, Serena Oliveri, Filippo Martinelli Boneschi, Alberto Priori

Submitted to: Journal of Medical Internet Research
on: February 27, 2025

Table of Contents

Is it time for the neurologist to use Large Language Models in everyday practice?

Natale Vincenzo Maiorana[1]; Sara Marceglia[1]; Mauro Treddenti[1]; Mattia Tosi[1]; Matteo Guidetti[1]; Maria Francesca Creta[1]; Tommaso Bocci[1]; Serena Oliveri[1]; Filippo Martinelli Boneschi[1]; Alberto Priori[1]

[1] University of Milan Milan IT

Corresponding Author:
Sara Marceglia

University of Milan
Via Antonio di Rudinì, 8
Milan
IT

Abstract

Background: Large Language Models (LLMs) such as ChatGPT and Gemini are increasingly explored for their potential in medical diagnostics, including neurology. Their real-world applicability remains inadequately assessed, particularly in clinical workflows where nuanced decision-making is required.

Objective: To evaluate the diagnostic accuracy and appropriateness of clinical recommendations provided by ChatGPT and Gemini compared to neurologists using real-world clinical cases.

Methods: This study consisted of a two-phase approach: (1) a systematic review of the literature on LLMs in neurology diagnosis to assess the adequacy of applied methodologies for clinical translation, and (2) an experimental evaluation of LLMs' diagnostic performance presenting real-world neurology cases to ChatGPT and Gemini, comparing their performance with that of clinical neurologists. The study was conducted simulating a first visit using information from anonymized patient records from the neurology department of the ASST Santi Paolo e Carlo Hospital (Milan, Italy), ensuring a real-world clinical context. A cohort of 28 anonymized patient cases was selected based on routine neurology consultations. These cases covered a range of neurological conditions and diagnostic complexities representative of daily clinical practice. The primary outcome was diagnostic accuracy of both neurologists and LLMs, defined as concordance with discharge diagnoses. Secondary outcomes included the appropriateness of recommended diagnostic tests and the extent of additional prompting required for accurate responses.

Results: Among the 24 studies identified in the literature review, most exhibited heterogeneous methodologies with structured prompts, specifically designed for the interaction with LLMs, but lacked real-world case evaluations. In the experimental phase, neurologists achieved a diagnostic accuracy of 75%, outperforming ChatGPT (54%) and Gemini (46%). Both LLMs demonstrated limitations in nuanced clinical reasoning and over-prescribed diagnostic tests in 17–25% of cases. Additionally, complex or ambiguous cases required further prompting to refine AI-generated responses.

Conclusions: While LLMs show potential as supportive tools in neurology, they currently lack the depth required for independent clinical decision-making. Future research should focus on refining LLM capabilities and developing evaluation methodologies that reflect the complexities of real-world neurological practice, thus ensuring effective, responsible, and safe use of such promising technologies.

(JMIR Preprints 27/02/2025:73212)
DOI: https://doi.org/10.2196/preprints.73212

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
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Is it time for the neurologist to use Large Language Models in everyday practice?

Dr. Natale Vincenzo Maiorana PhD[1], Dr. Sara Marceglia PhD[1], Dr. Mauro Treddenti[1,2], Dr. Mattia Tosi[1,2], Dr. Matteo Guidetti PhD[1], Dr. Maria Francesca Creta[1,2], Dr. Tommaso Bocci M.D.[1,2], Dr. Serena Oliveri PhD[1,2], Dr. Filippo Martinelli-Boneschi M.D. PhD[1,2], Dr. Alberto Priori M.D. PhD[1,2]

1 "Aldo Ravelli" Center for Neurotechnology and Experimental Brain Therapeutics, Department of Health Sciences, University of Milan, Via Antonio di Rudinì 8, 20142 Milan, Italy
2 Clinical Neurology Unit, "Azienda Socio-Sanitaria Territoriale Santi Paolo e Carlo", Department of Health Sciences, University of Milan, Via Antonio di Rudinì 8, 20142 Milan, Italy

Corresponding Author:
Dr. Sara Marceglia Ph.D.
Address: Department of Health Sciences, University of Milan, Via Antonio di Rudinì 8, 20142 Milan, Italy
Email: sara.marceglia@unimi.it
Phone Number: +39 02 50323233

Abstract

Background: Large Language Models (LLMs) such as ChatGPT and Gemini are increasingly explored for their potential in medical diagnostics, including neurology. Their real-world applicability remains inadequately assessed, particularly in clinical workflows where nuanced decision-making is required.

Objective: To evaluate the diagnostic accuracy and appropriateness of clinical recommendations provided by ChatGPT and Gemini compared to neurologists using real-world clinical cases.

Methods: This study consisted of a two-phase approach: (1) a systematic review of the literature on LLMs in neurology diagnosis to assess the adequacy of applied methodologies for clinical translation, and (2) an experimental evaluation of LLMs' diagnostic performance presenting real-world neurology cases to ChatGPT and Gemini, comparing their performance with that of clinical neurologists. The study was conducted simulating a first visit using information from anonymized patient records from the neurology department of the ASST Santi Paolo e Carlo Hospital (Milan, Italy), ensuring a real-world clinical context. A cohort of 28 anonymized patient cases was selected based on routine neurology consultations. These cases covered a range of neurological conditions and diagnostic complexities representative of daily clinical practice. The primary outcome was diagnostic accuracy of both neurologists and LLMs, defined as concordance with discharge diagnoses. Secondary outcomes included the appropriateness of recommended diagnostic tests and the extent of additional prompting required for accurate responses.

Results: Among the 24 studies identified in the literature review, most exhibited heterogeneous methodologies with structured prompts, specifically designed for the interaction with LLMs, but lacked real-world case evaluations. In the experimental phase, neurologists achieved a diagnostic accuracy of 75%, outperforming ChatGPT (54%) and Gemini (46%). Both LLMs demonstrated limitations in nuanced clinical reasoning and over-prescribed diagnostic tests in 17–25% of cases. Additionally, complex or ambiguous cases required further prompting to refine AI-generated responses.

Conclusions: While LLMs show potential as supportive tools in neurology, they currently lack the depth required for independent clinical decision-making. Future research should focus on refining LLM capabilities and developing evaluation methodologies that reflect the complexities of real-world neurological practice, thus ensuring effective, responsible, and safe use of such promising technologies.

Keywords:
Neurology; clinical practice; artificial intelligence; large language model, chat-gpt, gemini

Introduction

In recent years, large language models (LLM) and Generative Pre-trained Transformer (GPT) implementations opened the way for a fluid and natural human robot interaction [1], with the expectation to further enhance the revolution expected from artificial intelligence massive application, also in the healthacare field [1]

LLMs understand and generate human-like text [2], and are trained on massive datasets, enabling them to perform tasks like summarization, translation, and conversation [3].

Consumer available LLMs, such as Chat-GPT made by OpenAI and Gemini made by Google, enable users to easily interact enhancing accessibility for both personal and professional purposes.

Physicians might receive rapid support in analyzing clinical data, interpreting results, and formulating hypotheses [4], thus supporting decision making (diagnosis, treatment) [5] as well as education [6].

Ethical considerations and potential risks are however around the corner [7].

As a first point, the extensive, but generalist, datasets used for training may limit their performance in specialized domains [8] and may generate an inherent potential for bias[9]. To date, there is a lack of control and transparency on the data used to train the models[10] thus influencing the accuracy and reliability of the outputs [11].

Another point regards the excessive reliance on these tools that could lead to reduced human oversight and critical thinking which are of fundamental importance for medical decision-making [12].

Additionally, concern resides in the risk that patients with direct access to these models may misinterpret symptoms or engage in self-diagnosis, potentially leading to negative health consequences [13]. Responsible usage, therefore, must be promoted among healthcare professionals and patients, emphasizing a balanced integration between ai powered LLMs and humans agents [11].

On the other hand ,recent studies highlighted the increasing role of LLM-GPTs in the field of neurology, showing promising applications and fast progression across diagnostic and evaluative tasks [14]. For instance, ChatGPT was evaluated on neurology board-style exams, and surpassed the human average, showing strength in both lower-order and higher-order reasoning tasks, suggesting potential uses in neurology education and diagnostic support [15].

When evaluating LLMs in the field of neurology, it is crucial to consider that LLMs can be guided using different types of prompts, such as zero-shot prompts, which provide no prior task-specific guides, and few-shot prompts, including examples to help the model understand the request[18]. Structured prompts are highly detailed, specifying the context, the role of the model, and the desired output format, while iterative prompts involve a collaborative refinement process between the user and the model to optimize clarity and effectiveness[18].

The format of input questions or text passed to the model—whether open-ended or multiple-choice —also significantly impacts the accuracy of LLM responses [19]. Open-ended questions allow for a wide range of responses, which can lead to variability in accuracy depending on how well the prompt guides the model. Multiple-choice formats, however, constrain the model to select from predefined options, potentially improving accuracy by limiting the response scope[19].

However, the existing literature on clinical decision-making tools and physician-assistive technologies mostly relies on carefully curated scenarios and inputs designed explicitly for research purposes (e.g., educational vignettes or single cases reported in the literature) [16,17].
These curated prompts and sanitized datasets differ significantly from the high-pressure, overcrowded clinical settings where physicians make rapid decisions under suboptimal conditions. How do these tools perform in the unstructured, fragmented reality of clinical practice? Specifically, when used by fatigued clinicians facing systemic inefficiencies and competing priorities, does their effectiveness decline? Are there associated risks?
These questions can be answered only by assessing the diagnostic accuracy and reliability of LLMs in real-world applications, using data not specifically prepared, and without a structured interaction workflow.
In the present study, we focused on Gemini (Google) and ChatGPT (OpenAI) as the most accessible LLMs, then reviewed the literature to define their assessment methods. Next, we conducted an experiment comparing them to neurologists on real cases, evaluating their diagnostic accuracy and test recommendations. This simulation does not aim to assess system performance under ideal conditions but rather to examine how these tools integrate into real-world, imperfect settings. Our goal is to uncover critical insights into their usability, reliability, and potential pitfalls when deployed in their intended environments. To this end, we introduce the "Lazy Doctor" paradigm, where LLMs are used in their rawest form to simulate the worst possible misuse scenarios.

Methods

Literature review

We performed a literature search on PubMed (on December 1st, 2024), using the following search string: ((chatgpt) OR (chat-gpt) OR (gpt)) OR (gemini) AND neurology. The search was restricted to studies published between January 1st 2019 to December 1st 2024 (date of the search). Studies were included in the review if they utilized LLMs for diagnostic purposes in neurology or where their knowledge was assessed via neurology related questions. A total of 201 references were initially identified. The final sample consisted of 24 studies [14,15,20–41] (Table 1), which were included for data extraction. Results were categorized according to the type of prompt used and to the type of use-case. We considered zero-shot and few-shot examples of *soft prompting*, while structured and iterative prompts are categorized as *hard prompting*, given their high degree of detail and systematic design to guide the model's responses effectively. Furthermore, studies in which questions were presented in multiple-choice format were considered to be hard prompted, as multiple-choice questions create a context that influences the interpretation of a specific request by the user. The types of cases used in the studies were recorded and categorized into real cases, simulated cases, and questions, based on the specific material provided to the LLMs.

Experiment

A total of 56 consecutive patients admitted to the University Hospital Santi Paolo e Carlo -

Neurology department (Milan) between 1st July and 31th August 2023 were analyzed. Inclusion criteria were that cases should refer to patients who were admitted by a neurologist with a suspected diagnosis, cases should include a discharge letter containing the investigations conducted during the admission and a definitive diagnosis. Patients whose admission was purely therapeutic or observational, without a diagnostic purpose, were excluded. Additionally, patients with an already-known diagnosis were excluded.

The present study qualifies as a non-interventional observational study and does not constitute a clinical trial. The study has been approved by the IRB of the University of Milan with the approval number 123/24, All. 3 CE 10/12/24.

Patients provided consent for the processing of their personal data at the time of hospital admission under protocol ast_daz_502_ed00, in accordance with privacy regulations (D.Lgs. 101/2018, implementing EU GDPR 2016/679). This consent permits research on clinical data. The data have been anonymized prior to processing, ensuring that they cannot be traced back to any specific patient.

Simulation scenario

We retrieved the Electronic Health Records (EHRs) of the selected patients (pdf format) as they were filled in at the time of the patient's admission. We simulated the process of the initial diagnosis considering the procedure depicted in Figure 1. The patient, coming from the ER, is admitted to the Neurology ward. The neurologist in charge of admitting the patient begins the examination by collecting family history, history of present illness, past medical history, and reviewing the diagnostic results performed in the Emergency Department (ER). Additionally, the neurologist examines lab analysis results when available. At this point, the neurologist formulates a first diagnostic hypothesis that is recorded as the diagnosis at admission.

For each selected patient, the clinical data available to the neurologist at the time of admission were collected. Not all EHRs were structured identically. Some patients had more detailed clinical or family histories or underwent different clinical examinations in the ER.

Data presentation

Clinical cases were presented to ChatGPT-3.5 and Gemini, both in their free versions, in September 2024, to assess their reliability in providing neurological diagnoses.

The clinical cases were presented by providing the emergency EHR, present and past medical, familiar and social history, which referred patients to the neurology unit for further evaluation. This ensured that both ChatGPT and Gemini received the same information available to neurologists at the time of admission in a raw non-formatted form.

EHRs, present and past medical, familiar and social history were directly copied and pasted into the dialogue box of the LLMs after being fully anonymized, ensuring compliance with privacy regulations.

After the first patient case, subsequent cases were sequentially added in the same session. This methodology aimed to replicate a typical use case in which a physician interacts with the model pragmatically, focusing on obtaining responses rather than ensuring the model's precise understanding of the input data.

Cases were presented by an experimenter unaware of the final diagnosis made by the neurologist to

both models under standardized conditions, using the following prompts:

- ChatGPT prompt: *"Now I will give you a clinical case of a hypothetical patient. You will need to indicate the most likely diagnosis and suggest a series of clinical tests to confirm or rule out the diagnosis. Is that okay?"*
- Gemini prompt: *"I'm going to give you some made-up neurological cases. Your job is to tell me what you think is wrong with the patient and what tests you'd order."*

The prompts aimed to bypass safeguards preventing harmful diagnostic or treatment suggestions, without specifying cases, formatting, or response structure.

Outcomes and data analysis

The diagnostic accuracy for the neurologists, ChatGPT, and Gemini was evaluated based on their ability to match the patient's discharge diagnosis. The diagnostic accuracy of the neurologists was evaluated comparing the suspected diagnosis recorded at the admission to the neurology ward to the discharge diagnosis.

The evaluation considered also the agreement among neurologists, ChatGPT, and Gemini. Additionally, the number of instances in which both LLMs recommended all necessary tests, partially recommended them, or overprescribed unnecessary tests was recorded. Throughout the process, instances where LLMs required additional information or failed to provide a valid response were also documented.

Results

Literature review

Three main case types were identified: standardized questions were used in 11 studies, real cases in 10 studies, and simulated cases in 3 studies. Prompt usage varied significantly across studies. Hard prompts were the most frequently employed, appearing in 18 studies, seven studies involving real cases and three involving simulated cases. Soft prompts were used in five studies, of which only two assessing models' accuracy with real cases. Only two studies combined a soft prompt approach, open ended question and real cases to compare the accuracy of at least two LLMs. The best performance was recorded by ChatGPT-4 in a study conducted by Abbas et al. (2024) [34], where it achieved 100% accuracy in diagnostic questions, using structured prompts and multiple-choice. Conversely, the worst performance was observed in a study by Finelli (2024) [21], which employed open-ended prompts and real-case inputs. ChatGPT-4 failed to provide any correct diagnoses, while Glass AI achieved only one correct diagnosis.

Experiment

Clinical case records from 56 patients were analyzed. Twenty-eight patients (Mean age: 58.2 years; sex: 16 females) who met the inclusion criteria were selected for the study.

Both ChatGPT and Gemini understood the prompt without requiring further specifications. ChatGPT responded only with the request of the first clinical case, Gemini responded with an explanation of how the future responses would be organized indicating that the scheme of the response would be composed by three sections: diagnostic hypothesis, differential diagnosis and recommended

diagnostic test.

After the first response, ChatGPT needed further indication in nine cases out of 28 to achieve complete responses. In one case (id = 7), the experimenter requested clarification to determine whether the condition was considered primarily psychiatric or neurological. In six cases (id = 10, 12, 15, 17, 22, 28) ChatGPT needed to be prompted two times to provide both the missed diagnosis and clinical tests. In two cases (id = 6, 18) specific clinical tests were solicited to confirm or exclude a diagnosis. Conversely, Gemini generally provided an organized response framework and required explicit prompting to deliver a diagnosis only in one case (id =28).

In none of the cases ChatGPT or Gemini showed hallucination defined as a nonfactual, nonsensical, or inconsistent response to a clear given prompt [42].

In relation to the accuracy of the proposed diagnosis, Neurologists correctly diagnosed 75 % of cases, while ChatGPT achieved an accuracy of 54 % and Gemini 46 %.

The diagnostic accuracy was evaluated across a range of neurological disorders, revealing notable differences in performance.

In infectious diseases, all three raters—neurologist, ChatGPT, and Gemini—demonstrated equal accuracy, correctly diagnosing 50% of the cases. For movement disorders, both neurologists and ChatGPT correctly diagnosed 75% of the cases, while Gemini's performance was notably lower, correctly diagnosing only 25%.

For neuromuscular disorders, neurologists, ChatGPT, and Gemini all displayed equal accuracy, correctly diagnosing 67 % of the cases, leaving 33 % of the cases misdiagnosed by each rater. In psychiatric disorders, neurologists and Gemini both achieved a 50% accuracy rate, whereas ChatGPT underperformed, correctly diagnosing only 25% of the cases.

In vascular disorders, neurologists and Gemini both had an accuracy of 75%, while ChatGPT performed slightly lower, with a 62 % correct diagnosis rate.

Overall, neurologists achieved the highest diagnostic accuracy across all pathologies, correctly diagnosing 75% of the cases. ChatGPT followed with an overall accuracy of 54%, and Gemini achieved an accuracy of 46% (Fig. 2).

In terms of diagnostic agreement, all three (neurologists, ChatGPT, and Gemini) agreed providing the correct diagnosis in 36 % of cases. In 25 % of cases, all three were incorrect. In 11% of cases only the neurologists were correct and both ChatGPT-3.5 and Gemini were incorrect. ChatGPTonly in the 11% of cases, while both the neurologist and Gemini were accurate, the percentage was 11 % of cases, and only Gemini was wrong in 18 % of cases.

In none of the cases ChatGPT and Gemini were correct while Neurologist failed in indicating the correct diagnosis (Fig. 3a).

Regarding the diagnostic test indication, both ChatGPT and Gemini recommended the correct set of basic tests in 55 % of cases. However, there was agreement between ChatGPT and Gemini in suggesting the correct diagnostic tests in only 32 % of cases (Fig. 3b).

In terms of over-prescription of diagnostic tests, ChatGPT recommended an excessive number of tests in 17 % of cases, while Gemini did so in 25%. In 64% of instances neither ChatGPT or Gemini overprescribed unnecessary tests. Gemini alone over-prescribed tests in 18% of cases, ChatGPT alone did so in 11%, and both models suggested excessive tests in 7% of cases.

Discussion

This study aimed to address the critical gap in real-world evidence regarding the use of LLMs in assisting daily clinical practice within a neurological setting. First, our literature review revealed a striking underutilization of soft prompts in real-case scenarios, despite their potential to better simulate the complexity of everyday clinical interactions. This gap underscored the need for a more ecologically valid evaluation of LLMs in diagnostic settings.

Second, to bridge this gap, we conducted an experiment designed to test LLM performance under real-world conditions. Our results, obtained from a cohort of patients with diverse neurological conditions, demonstrated that in non-structured environments, neurologists consistently outperformed LLMs in terms of diagnostic accuracy. While both neurologists and LLMs achieved comparable accuracy in certain disease categories, such as infectious diseases and neuromuscular disorders, neurologists demonstrated superior performance in movement disorders, psychiatric disorders, and vascular disorders. These conditions often require nuanced clinical judgment, experience in interpreting subtle clinical signs, and the ability to integrate information [43]. These results align with the broader literature on LLMs in specialized medical contexts, where human expertise surpasses LLMs' performance, especially in areas requiring nuanced clinical judgment, advanced interpretative skills, and integrative diagnostic approaches [44]. Regarding diagnostic agreement, there was substantial variability among the three raters. While all three agreed on the correct diagnosis in a significant proportion of cases, in no cases the wrong response was given by the neurologists only. Our findings align with studies showing that LLMs accuracy in clinical recommendation varies with the severity of the presentation and the complexity of the clinical picture: it performs well when initial symptoms clearly indicate a condition, but less so with more ambiguous cases [45]. LLMs appear accurate on the surface when managing well-known, sharply defined symptoms but often fail to provide appropriate clinical management recommendations [45]. This shortfall likely arises from the lack of deep, discipline-specific knowledge and of a clear understanding of clinical constraints, limiting their ability to navigate patient's management effectively, beyond initial diagnosis [45]. Both models, especially Gemini, showed test over-prescription, highlighting LLMs' inability to apply cost-benefit reasoning in clinical care.

While emphasizing the promise of LLMs in clinical practice, it should be noted their limitations in handling complex diagnosis and in recommendation of further diagnostic assessment.

LLMs can support clinicians in structured assessment contexts, but they are still limited in their ability to account for the subtle and often ambiguous nature of real-world clinical decision-making [43]. Therefore, the "lazy doctor" may use available LLMs as support, but, without additional specific training, this use would not increase diagnostic accuracy, even though not causing harms. The major risk at present is the over-prescription of recommended diagnostic tests that may impact on costs associated to the diagnostic pathway.

The ability of LLMs to adopt a human-like interaction pattern can be advantageous in human-machine interaction. However, this same feature also increases the risk of misunderstandings due to natural language, unlike typical clinical decision support systems, which rely on standardized and unambiguous communication protocols. The use of LLMs requires careful management to ensure consistency and prevent misinterpretation of the information they provide.

New studies should focus on how models are utilized by clinicians within a human-in-the-loop framework, testing how experts use LLMs. For example, a study by Goh and colleagues [46] showed that while LLMs alone performed better than clinicians using conventional resources, integrating

LLMs with expert oversight could enhance diagnostic accuracy and efficiency indicating the potential for significant improvements in clinical practice when AI tools are effectively integrated with human expertise.
Future works should focus on refining LLMs' algorithms to better interpret clinical data and test the reliability of responses using large datasets of real-world clinical cases, creating models that complement human expertise.

Funding: This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of Interest Disclosure: The authors declare no conflicts of interest

Abbreviations: LLM = Large Language Model; GPT = Generative Pre-training Transformer

Data availability: the data that support the findings of this study are available upon reasonable request

Bibliography

1. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. *Neurocomputing*. 2017;237:350-361. doi:10.1016/j.neucom.2017.01.026

2. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners.

3. Basyal L, Sanghvi M. Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. Published online October 17, 2023. doi:10.48550/arXiv.2310.10449

4. Zhang S, Song J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. *Sci Rep*. 2024;14(1):17118. doi:10.1038/s41598-024-67429-4

5. Goodman RS, Patrinely JR, Osterman T, Wheless L, Johnson DB. On the cusp: Considering the impact of artificial intelligence language models in healthcare. *Med*. 2023;4(3):139-140. doi:10.1016/j.medj.2023.02.008

6. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Medical Education*. 2024;58(11):1276-1285. doi:10.1111/medu.15402

7. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *Journal of Medical Internet Research*. 2023;25(1):e48568. doi:10.2196/48568

8. Rane N, Choudhary S, Rane J. Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation. Published online February 13, 2024.

doi:10.2139/ssrn.4723687

9. Shah SV. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. *JAMA Network Open*. 2024;7(8):e2425953. doi:10.1001/jamanetworkopen.2024.25953

10. Liesenfeld A, Lopez A, Dingemanse M. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In: *Proceedings of the 5th International Conference on Conversational User Interfaces*. CUI '23. Association for Computing Machinery; 2023:1-6. doi:10.1145/3571884.3604316

11. Frosolini A, Gennaro P, Cascino F, Gabriele G. In Reference to "Role of Chat GPT in Public Health", to Highlight the AI's Incorrect Reference Generation. *Ann Biomed Eng*. 2023;51(10):2120-2122. doi:10.1007/s10439-023-03248-4

12. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Research Square*. Published online February 28, 2023:rs.3.rs. doi:10.21203/rs.3.rs-2566942/v1

13. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wien Klin Wochenschr*. 2024;136(7-8):236-238. doi:10.1007/s00508-024-02329-1

14. Fonseca Â, Ferreira A, Ribeiro L, Moreira S, Duque C. Embracing the future-is artificial intelligence already better? A comparative study of artificial intelligence performance in diagnostic accuracy and decision-making. *Eur J Neurol*. 2024;31(4):e16195. doi:10.1111/ene.16195

15. Schubert MC, Wick W, Venkataramani V. Evaluating the Performance of Large Language Models on a Neurology Board-Style Examination. Published online July 14, 2023:2023.07.13.23292598. doi:10.1101/2023.07.13.23292598

16. Madadi Y, Delsoz M, Lao PA, et al. ChatGPT Assisting Diagnosis of Neuro-Ophthalmology Diseases Based on Case Reports. *Journal of Neuro-Ophthalmology*. Published online April 28, 2022:10.1097/WNO.0000000000002274. doi:10.1097/WNO.0000000000002274

17. Kozel G, Gurses ME, Gecici NN, et al. Chat-GPT on brain tumors: An examination of Artificial Intelligence/Machine Learning's ability to provide diagnoses and treatment plans for example neuro-oncology cases. *Clinical Neurology and Neurosurgery*. 2024;239:108238. doi:10.1016/j.clineuro.2024.108238

18. Heston TF, Khun C. Prompt Engineering in Medical Education. *International Medical Education*. 2023;2(3):198-205. doi:10.3390/ime2030019

19. Rodrigues L, Pereira FD, Cabral L, Gašević D, Ramalho G, Mello RF. Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence*. 2024;7:100248. doi:https://doi.org/10.1016/j.caeai.2024.100248

20. Altunisik E, Firat YE, Cengiz EK, Comruk GB. Artificial intelligence performance in clinical neurology queries: the ChatGPT model. *Neurol Res*. 2024;46(5):437-443.

doi:10.1080/01616412.2024.2334118

21. Finelli PF. Neurological Diagnosis: Artificial Intelligence Compared With Diagnostic Generator. *Neurologist*. 2024;29(3):143-145. doi:10.1097/NRL.0000000000000560

22. Patel MA, Villalobos F, Shan K, et al. Generative artificial intelligence versus clinicians: Who diagnoses multiple sclerosis faster and with greater accuracy? *Mult Scler Relat Disord*. 2024;90:105791. doi:10.1016/j.msard.2024.105791

23. Galetta K, Meltzer E. Does GPT-4 have neurophobia? Localization and diagnostic accuracy of an artificial intelligence-powered chatbot in clinical vignettes. *Journal of the Neurological Sciences*. 2023;453:120804. doi:10.1016/j.jns.2023.120804

24. Chen TC, Kaminski E, Koduri L, et al. Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales. *World Neurosurgery*. 2023;179:e342-e347. doi:10.1016/j.wneu.2023.08.088

25. Du X, Novoa-Laurentiev J, Plasek JM, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *eBioMedicine*. 2024;109:105401. doi:https://doi.org/10.1016/j.ebiom.2024.105401

26. Lin SY, Hsu YY, Ju SW, Yeh PC, Hsu WH, Kao CH. Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023. *Digit Health*. 2024;10:20552076241291404. doi:10.1177/20552076241291404

27. Wang X, Ye S, Feng J, Feng K, Yang H, Li H. Performance of ChatGPT on prehospital acute ischemic stroke and large vessel occlusion (LVO) stroke screening. *DIGITAL HEALTH*. 2024;10:20552076241297127. doi:10.1177/20552076241297127

28. Tailor PD, Dalvin LA, Starr MR, et al. A Comparative Study of Large Language Models, Human Experts, and Expert-Edited Large Language Models to Neuro-Ophthalmology Questions. *J Neuroophthalmol*. Published online April 2, 2024. doi:10.1097/WNO.0000000000002145

29. Panagiotis Giannos. Evaluating the limits of AI in medical specialisation: ChatGPTâ™s performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurology Open*. 2023;5(1):e000451. doi:10.1136/bmjno-2023-000451

30. Shukla R, Mishra AK, Banerjee N, Verma A. The Comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for Diagnosing Cases of Neuro-Ophthalmology. *Cureus*. 2024;16(4):e58232. doi:10.7759/cureus.58232

31. Tse Chiang Chen, Evan Multala, Patrick Kearns, et al. Assessment of ChatGPTâ™s performance on neurology written board examination questions. *BMJ Neurology Open*. 2023;5(2):e000530. doi:10.1136/bmjno-2023-000530

32. Williams SC, Starup-Hansen J, Funnell JP, et al. Can ChatGPT outperform a neurosurgical trainee? A prospective comparative study. *Br J Neurosurg*. Published online February 2, 2024:1-10. doi:10.1080/02688697.2024.2308222

33. Lee JH, Choi E, McDougal R, Lytton WW. GPT-4 Performance for Neurologic Localization. *Neur Clin Pract*. 2024;14(3):e200293. doi:10.1212/CPJ.0000000000200293

34. Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus*. 2024;16(3):e55991. doi:10.7759/cureus.55991

35. Hewitt KJ, Wiest IC, Carrero ZI, et al. Large language models as a diagnostic support tool in neuropathology. *J Pathol Clin Res*. 2024;10(6):e70009. doi:10.1002/2056-4538.70009

36. Haemmerli J, Sveikata L, Nouri A, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? *BMJ Health Care Inform*. 2023;30(1). doi:10.1136/bmjhci-2023-100775

37. Wang C, Liu S, Li A, Liu J. Text Dialogue Analysis for Primary Screening of Mild Cognitive Impairment: Development and Validation Study. *Journal of Medical Internet Research*. 2023;25(1):e51501. doi:10.2196/51501

38. Pedro T, Sousa JM, Fonseca L, et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *J Neurointerv Surg*. Published online March 7, 2024:jnis-2024-021556. doi:10.1136/jnis-2024-021556

39. Nógrádi B, Polgár TF, Meszlényi V, et al. ChatGPT M.D.: Is there any room for generative AI in neurology? *PLoS One*. 2024;19(10):e0310028. doi:10.1371/journal.pone.0310028

40. Ros-Arlanzón P, Perez-Sempere A. Evaluating AI Competence in Specialized Medicine: Comparative Analysis of ChatGPT and Neurologists in a Neurology Specialist Examination in Spain. *JMIR Med Educ*. 2024;10:e56762. doi:10.2196/56762

41. Erdogan M. Evaluation of responses of the large language model GPT to the neurology question of the week. *Neurol Sci*. 2024;45(9):4605-4606. doi:10.1007/s10072-024-07580-y

42. Waldo J, Boussard S. GPTs and Hallucination: Why do large language models hallucinate? *Queue*. 2024;22(4):Pages 10:19-Pages 10:33. doi:10.1145/3688007

43. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

44. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? *Am J Emerg Med*. 2024;79:44-47. doi:10.1016/j.ajem.2024.02.008

45. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research*. 2023;25(1):e48659. doi:10.2196/48659

46. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A

Randomized Clinical Trial. *JAMA Network Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

Figures

Figure 1.
Procedure used to compare diagnosis and clinical tests assessment from neurologists and LLMs.

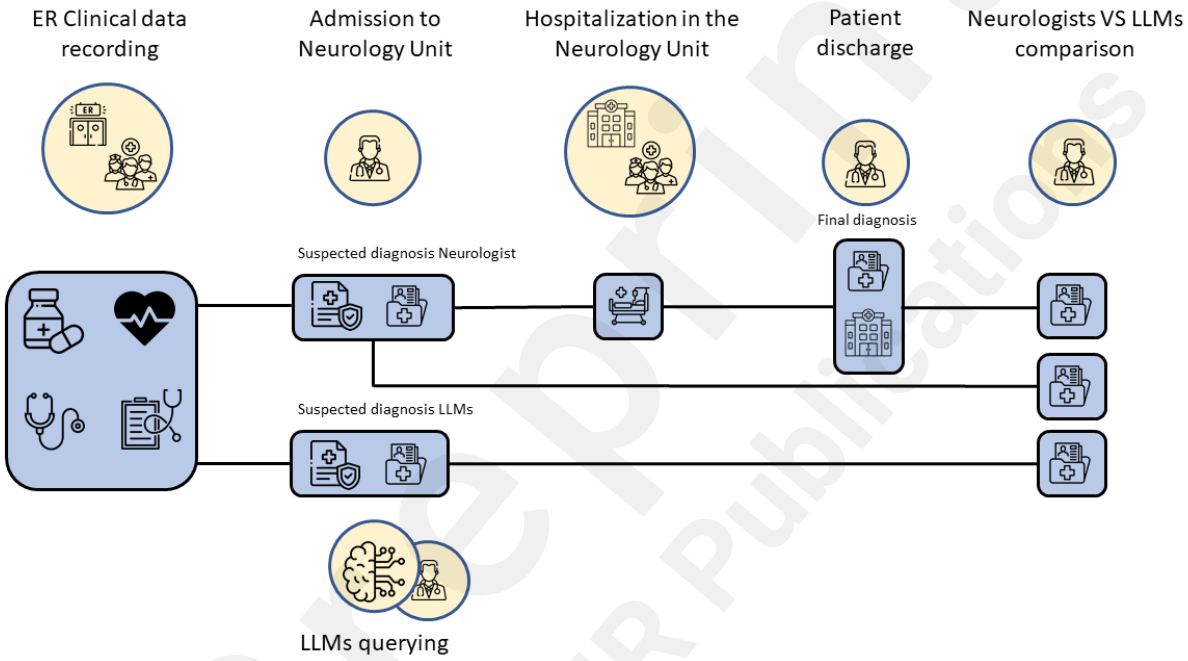


Figure 2.
Diagnostic accuracy of Neurologists, ChatGPT, and Gemini across neurological disease categories. Results are presented as percentages of the total cases (n = 28).

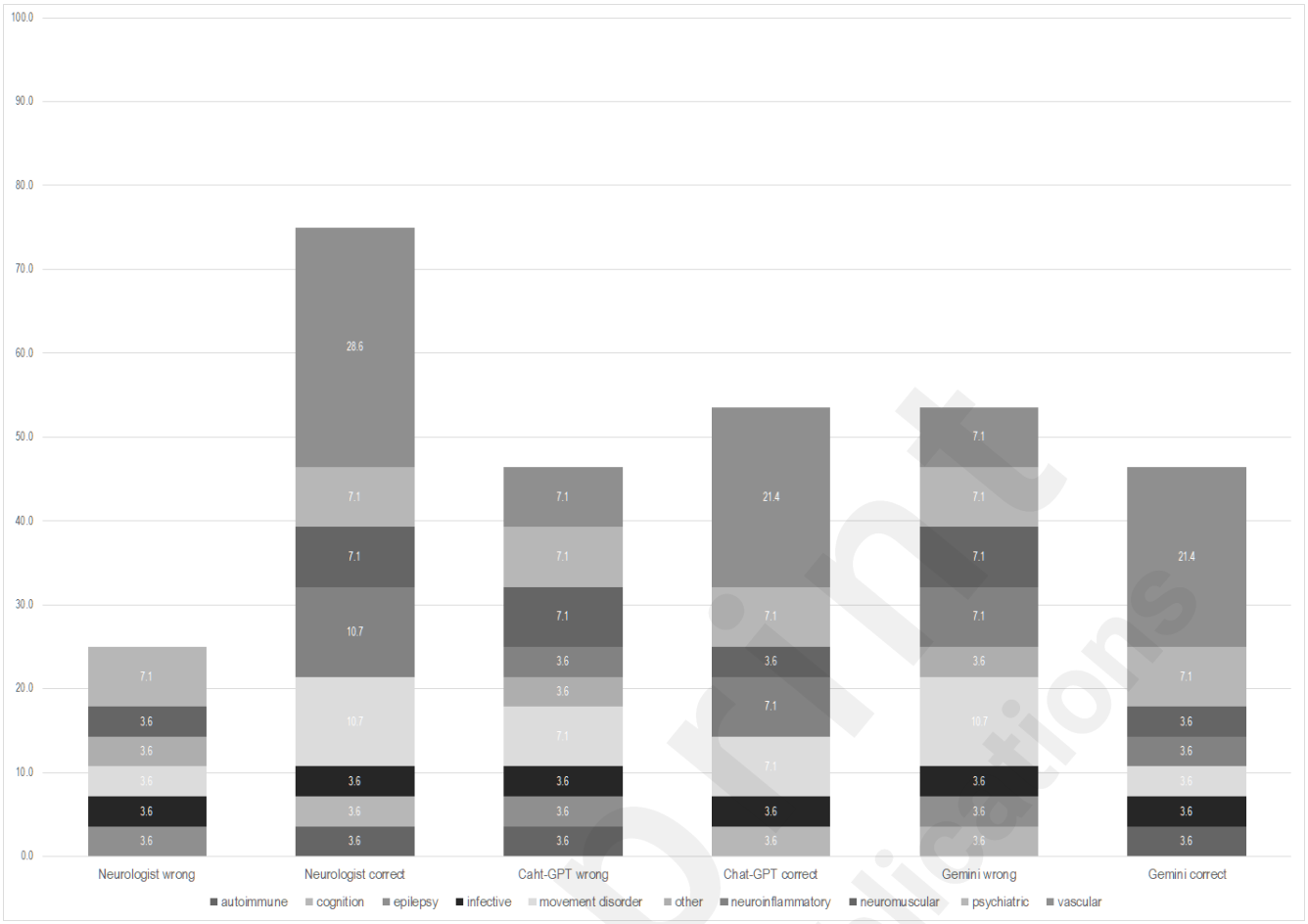
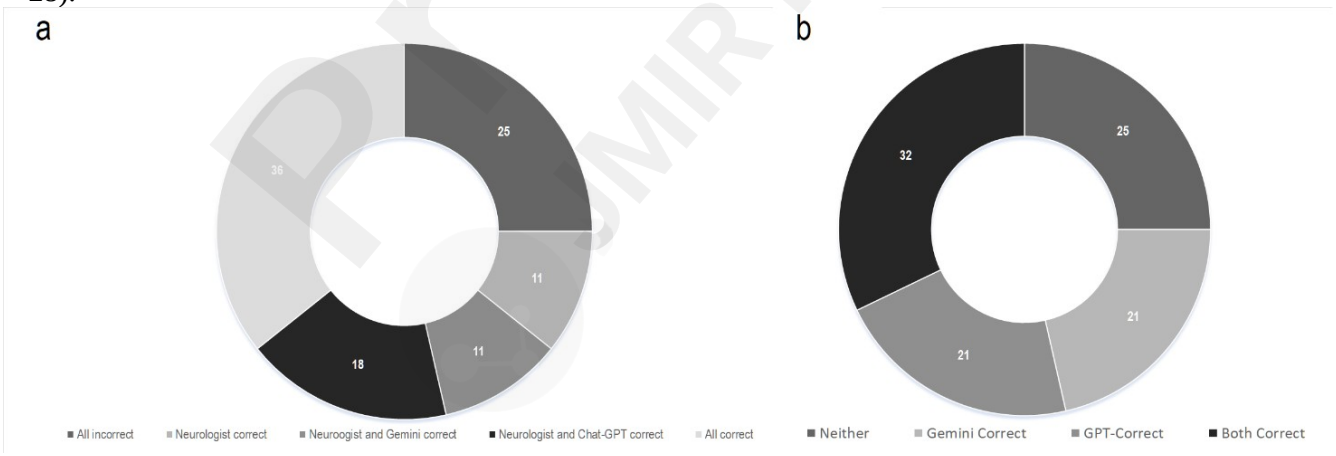


Figure 3.

a) Agreement between Neurologist, ChatGPT, and Gemini in indicating correct versus incorrect diagnoses. Results are presented as percentages of total cases (n = 28). b) Agreement between ChatGPT, and Gemini in recommending correct versus incorrect diagnostic tests. Results are presented as percentages of total cases (n = 28).



Tables

Table 1. Extracted and analyzed studies

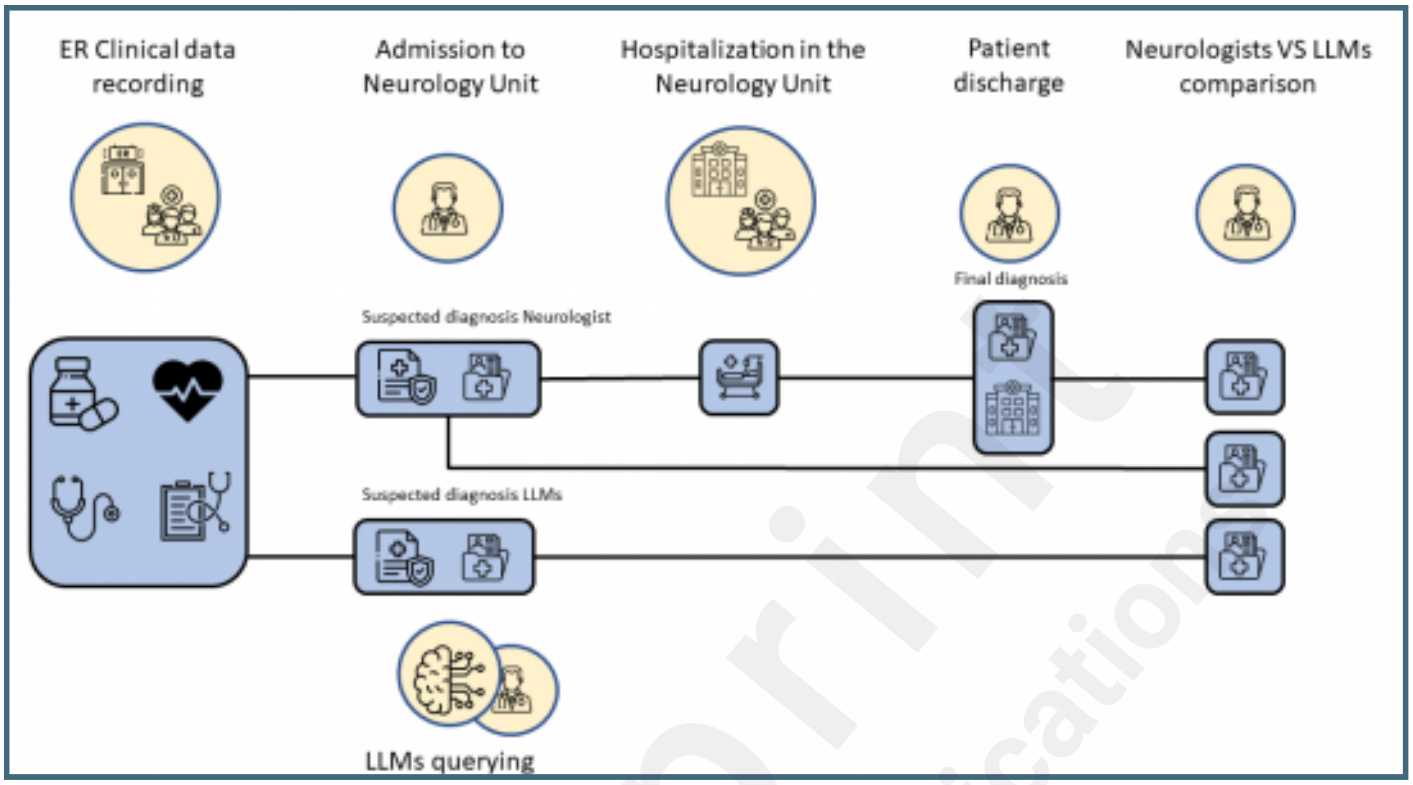
N	Authors	Year	Type of Prompt	Type of Input	Type of Cases	LLM	N. cases	Main Findings
1	Kristin Galetta, Ethan Meltzer [23]	2023	Hard	Open Ended	Simulated cases	GPT-4	29	GPT-4 correctly identified the lesion location in less than 50% of cases. GPT-4 showed higher accuracy in simpler cases compared to more complex ones.
2	Chen et al. [24]	2023	Hard	Open Ended	Simulated cases	GPT-4	20	ChatGPT demonstrated the ability to evaluate neurological exams using established scales but exhibited limitations in accuracy, especially with complex or vague descriptions. It showed an average error rate of 10% for basic GCS cases, with the error rate rising 45% for more complex cases.
3	Patel et al. [22]	2024	Hard	Open Ended	Real cases	GPT-3.5	100	ChatGPT-3.5 diagnosed MS faster than clinicians (0.08 vs. 0.35 years). After including MRI data, accuracy varied in function of patients' age, gender and ethnicity.
4	Du et al. [25]	2024	Hard	Open Ended	Real cases	GPT-4 Llama 2	4949 note from 1969 patients	GPT-4 demonstrated superior precision and efficiency compared to Llama 2. Best performance with 90.2% precision, 94.2% recall, and a 92.1% F1-score.
5	Lin et al. [26]	2024	Hard	Multiple choice	Questions	GPT-4o, Claude_3.5 Sonnet, Gemini Advanced	680	Claude 3.5 Sonnet answered 17 correctly (73.91%), GPT-4o answered 15 correctly (65.22%), and Gemini Advanced answered 11 correctly (47.83%).
6	Wang et al. [27]	2024	Hard	Open Ended	Real cases	GPT-3.5, GPT-4	400 records of 400 patients	GPT-4 outperformed GPT-3.5 in Acute Ischemic Stroke (AIS) screening (AUC 0.75 vs. 0.59) and Large Vessel Occlusion (LVO) identification (AUC 0.71 vs. 0.60).
7	Tailor et al. [28]	2024	Soft	Open Ended	Questions	GPT-3.5, GPT-4, Claude 2, Bing, Bard	21	Expert-edited LLM responses (Expert + AI) had the highest expert-determined ratings of quality and empathy. Expert + AI performed better than Expert responses in both quality and empathy.
8	Giannos [29]	2023	Hard	Multiple choice	Questions	GPT-3.5 Legacy, GPT-3.5 Default,	69	ChatGPT-3.5 failed to meet the 58% passing threshold for the 2022 SCE neurology exam. ChatGPT-4, achieved the highest accuracy of 64%.

Supplementary Files

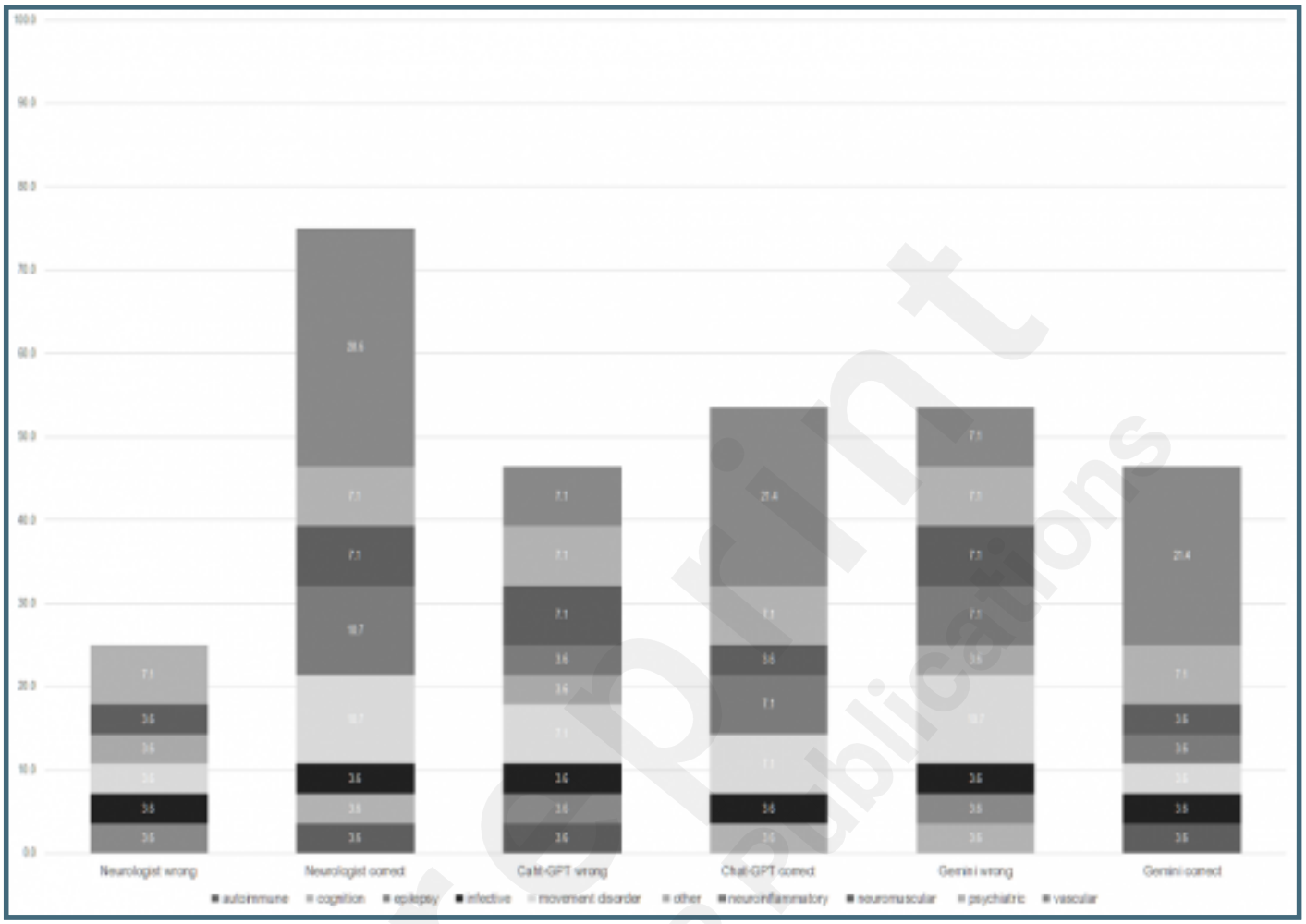
Untitled.
URL: http://asset.jmir.pub/assets/d33ce53fde9a54afe85a5ed289765eea.docx

Figures

Procedure used to compare diagnosis and clinical tests assessment from neurologists and LLMs.



Diagnostic accuracy of Neurologists, ChatGPT, and Gemini across neurological disease categories. Results are presented as percentages of the total cases (n = 28).



a) Agreement between Neurologist, ChatGPT, and Gemini in indicating correct versus incorrect diagnoses. Results are presented as percentages of total cases (n = 28). b) Agreement between ChatGPT, and Gemini in recommending correct versus incorrect diagnostic tests. Results are presented as percentages of total cases (n = 28).