NeuroRegulation

ISNR INTERNATIONAL SOCIETY FOR NEUROREGULATION & RESEARCH

Neurological Diagnosis in the AI Era: A Comparative Assessment of ChatGPT 3.5, Google Gemini, Being AI, and Perplexity AI

Muhammad Essa[1], Abdelrhman H. Mohamed[2], Zainullah[3], and Milica Jovanovic[4]*

[1]Neurology Department, BMC Hospital, Quetta, Pakistan
[2]Faculty of Medicine, Luxor University, Luxor, Egypt
[3]Sapienza University of Rome, Rome, Italy
[4]University of Tartu, Tartu, Estonia

Abstract

Background. In neurological diagnostics, where complexity, data volume, and diagnostic urgency present major obstacles, artificial intelligence (AI) systems have the potential to revolutionize the field. Despite widespread use, there is a lack of comparable performance assessments of publicly available AI tools for integrated clinical reasoning in neurology. **Methods.** This cross-sectional study evaluated five AI platforms (ChatGPT 3.5, Google Gemini, Bing AI, Perplexity AI, DeepSeek) utilizing 15 standardized neurological cases from *Case Files: Neurology, Third Edition*. Each platform was given identical prompts imitating clinical consultations. Responses were evaluated (maximum 6 points per case; total 90) in three domains: diagnosis, subsequent diagnostic step, and therapeutic/molecular foundation. Nonparametric statistical methods (Kruskal-Wallis, Chi-square) assessed performance disparities. **Results.** ChatGPT achieved the highest overall score (88/90, 97.8%), followed by DeepSeek (86/90, 95.6%), Perplexity (84/90, 93.3%), Google Gemini (78/90, 86.7%), and Microsoft Copilot (73/90, 81.1%). Therapeutic accuracy was 100% for ChatGPT, DeepSeek, and Gemini, whereas it was 80% for Copilot. Although there were disparities in performance, inferential statistics revealed no significant differences between platforms (Kruskal-Wallis p = .423; Chi-square p = .374). Verbosity showed significant variation: DeepSeek averaged 488 words per response, whereas Copilot and Perplexity averaged 239 to 240 words. **Conclusion.** Popular AI platforms (ChatGPT, DeepSeek) exhibit significant proficiency in neurological diagnosis and treatment planning, but there is a huge difference in the depth and structure of responses across all of the tools. AI should be used as complementary healthcare assistance, with future integration requiring better explainability and real-world validation.

Keywords: artificial intelligence; neurological diagnosis; large language models (LLMs); comparative evaluation; diagnostic accuracy (MeSH)

Citation: Essa, M., Mohamed, A. H., Zainullah, & Jovanovic, M. (2026). Neurological diagnosis in the AI era: A comparative assessment of ChatGPT 3.5, Google Gemini, Being AI, and Perplexity AI. *NeuroRegulation, 13*(1), 54–64. https://doi.org/10.15540/nr.13.1.54

***Address correspondence to**: Dr. Syed Muhammad Essa BMCH Complex. Brewery Road. Quetta, Balochistan. Pakistan, 87300. Email: dressakhan777@gmail.com

Edited by:
Rex L. Cannon, PhD, Currents, Knoxville, Tennessee, USA

Reviewed by:
Rex L. Cannon, PhD, Currents, Knoxville, Tennessee, USA
Randall Lyle, PhD, Mount Mercy University, Cedar Rapids, Iowa, USA

Introduction

In recent years, incorporating artificial intelligence (AI) into different sectors has transformed traditional problem-solving and decision-making methodologies (Nguyen & Vo, 2024; Shokran et al., 2025). AI has shown great promise in helping medical professionals with activities ranging from diagnosis to treatment planning, especially in the field of healthcare (Rashid & Sharma, 2025; Salammagari & Srivastava, 2024; Zeb et al., 2024). The development of more sophisticated AI has the potential to greatly improve neurological care (AbuAlrob & Mesraoua, 2024; Kalani & Anjankar, 2024). This subject is distinguished by its intricacy and dependence on the precise interpretation of complex data. Global healthcare systems have significant challenges with neurological disorders,

including stroke, epilepsy, Parkinson's disease, and Alzheimer's disease (Kandel, 2025; Yang et al., 2025). These conditions are frequently present with a wide range of symptoms and fluctuating rates of progression, which makes diagnosis and treatment extremely difficult. Conventional diagnostic approaches mostly depend on laboratory testing, neuroimaging technologies, and clinical experience, all of which can be laborious and subject to subjective interpretation.

By combining the strength of neural networks, machine learning algorithms, and big data analytics, AI technologies present a paradigm change in neurological diagnostics (Dipankar et al., 2025; Onciul et al., 2025). In order to help doctors diagnose patients accurately and quickly, AI systems have the capacity to evaluate enormous volumes of patient data, including genetic data, clinical records, and medical imaging (Alhejaily, 2025; Li et al., 2024; Oyeniyi & Oluwaseyi, 2024). AI has the potential to enhance human knowledge of neurological diseases by providing machine-driven insights, hence minimizing diagnosis errors and enhancing treatment techniques (Mennella et al., 2024; Valerio et al., 2025).

To determine their practical efficacy and therapeutic value, however, a thorough assessment and comparison are still required amid the expanding field of AI-powered diagnostic tools. This study aims to fill that gap by undertaking a comparative analysis of five major AI platforms in neurological diagnosis. We will specifically assess how well they perform in relation to important parameters like interpretability, scalability, sensitivity, specificity, and diagnostic accuracy. The comparative evaluation also helps researchers and medical professionals find the advantages and disadvantages of various AI systems and decide which ones are most appropriate for particular clinical circumstances or diagnostic tasks. Through meticulous comparison studies, scientists may offer important insights into how AI systems function in various scenarios, supporting the integration of AI technology into clinical practice and supporting evidence-based decision-making. Through thoroughly analyzing each system's advantages and disadvantages, we hope to offer insightful information that can guide clinical judgment and propel neurology forward.

Methodology

This was a cross-sectional comparative study evaluating the diagnostic performance of five publicly available AI platforms—ChatGPT (OpenAI), Microsoft Copilot (Microsoft), DeepSeek, Google Gemini (Google), and Perplexity—in interpreting standardized neurological case scenarios. All responses were generated in April–May 2024.

The four chosen AI platforms—ChatGPT 3.5, Google Gemini, Bing AI, and Perplexity AI—were presented with each case scenario as part of the study design. The following was the typical input prompt for every case: “I'm working on a neurology quiz and will provide you the patient's full medical history, presenting problems, and findings. Behaving like a neurology professor, please provide the most appropriate answer.”

Using a consistent input prompt, the study design presented case scenarios to the AI platforms in an organized manner. This prompt was designed to mimic an actual clinical setting, supplying pertinent findings, the patient’s medical history, and their current problems. Three key questions were used to help AI systems in the task of producing diagnosis theories and therapy recommendations:

1. A diagnosis?
2. What is the next diagnostic action?
3. Next therapeutic step or molecular/genetic basis, ordered by probability.

All AI platforms’ responses were carefully documented and assessed to make sure they answered the questions correctly and comprehensively. Crucially, the AI systems were impartially evaluated for their effectiveness in neurological diagnosis because they were not pretrained with particular command sets or queries.

Ethical Considerations

The study did not require formal ethics committee permission because it was conducted solely using published cases, did not include human individuals, and did not access personal health information. The use of anonymized data and standardized case scenarios guaranteed adherence to ethical rules regarding patient privacy and confidentiality.

Selection of Cases

To undertake a thorough comparative evaluation of AI systems in neurological diagnosis, 15 clinically significant neurological cases were chosen from *Case Files: Neurology, Third Edition* by Eugene C. Toy et al. (2018). These scenarios were selected to

represent a broad spectrum of neurological conditions, such as multiple sclerosis, epilepsy, stroke, Parkinson's disease, Alzheimer's disease, and traumatic brain injury. Every case was thoroughly examined to make sure it depicted a situation that was clinically realistic and had unique diagnostic issues and considerations.

LLM Input-Output Procedures

Every AI platform was given the same input prompt in the form of a chat session for every case scenario. The AI systems were able to produce responses that were well-informed since this input prompt supplied adequate clinical background. Because the AI platforms weren't pretrained with any particular command sets or questions, their responses could be evaluated objectively because they were based only on their natural capabilities.

Scoring System

Accuracy and comprehensiveness in answering the three main questions in the input prompt determined the score for each AI platform's response. This was the scoring system:

(a) 6 points (*excellent*): When all three questions—the diagnosis, the next diagnostic step, the molecular/genetic basis, or the next therapeutic step—were answered correctly.
(b) 4 points (*good*): When two questions were answered accurately.
(c) 2 points (*moderate*): When just one question was answered accurately.
(d) 1 point (*very poor*): When every question was answered wrong.

This grading system allows for a sophisticated evaluation of the AI platforms' performance, taking into account both the accuracy of diagnoses and the comprehensiveness of their responses.

Statistical Analysis

Descriptive statistics were computed for total scores, mean word count, and individual domain accuracy. To assess normality, both Kolmogorov-Smirnov and Shapiro-Wilk tests were applied. Given the nonnormal distribution of scores (as evidenced by significant *p*-values in both tests), nonparametric methods were employed.

Differences in overall performance across AI platforms were analyzed using the Kruskal-Wallis H test, followed by post hoc pairwise comparisons with Bonferroni correction for multiple testing. Word count differences were described descriptively. Exploratory subgroup analyses were planned to evaluate:

- Performance variation by diagnostic category (e.g., movement disorders vs. infectious diseases)
- Correlation between word count and accuracy
- Consistency of responses across repeated prompts

However, due to limited sample size and variability in output structure, these analyses were deemed exploratory and not formally reported in the final results.

Limitations of the Study Design

While efforts were made to maintain standardization, several limitations should be acknowledged:

- AI outputs were influenced by version updates and backend changes during the study period.
- Some models exhibited variability when prompted multiple times, limiting reproducibility.
- No formal external validation dataset was used.
- Human-level comparison (e.g., resident vs. AI) was not included but is recommended for future work.

Results

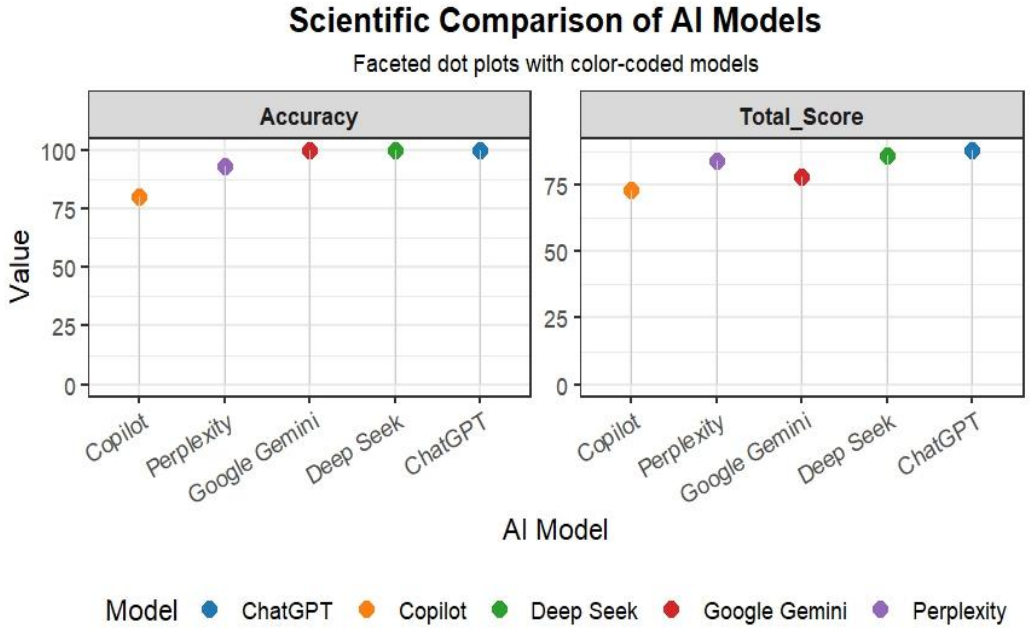
The basic assessment of AI platform performance was based on a detailed scoring system, with each of the 15 scenarios worth a maximum of 6 points, for a total possible score of 90 points. A thorough synopsis of each AI model's performance metrics is given in Table 1. ChatGPT achieved the highest overall score of 88 points, corresponding to an impressive 97.77% accuracy. DeepSeek closely behind with 86 points (95.55%), signifying strong performance. Perplexity attained third place with 84 points (93.33%). Google Gemini achieved 78 points (86.66%), and Copilot obtained the lowest overall score of 73 points (81.11%; As illustrated in Figure 1).

All AI models exhibited a steady median score of 6.00. This indicates that in at least 50% of the assessed cases, each AI platform was able to deliver an impeccable response by precisely resolving all three essential enquiries: diagnosis, subsequent diagnostic action, and the ensuing therapeutic/molecular/genetic rationale.

Table 1
Performance Metrics by AI Platform

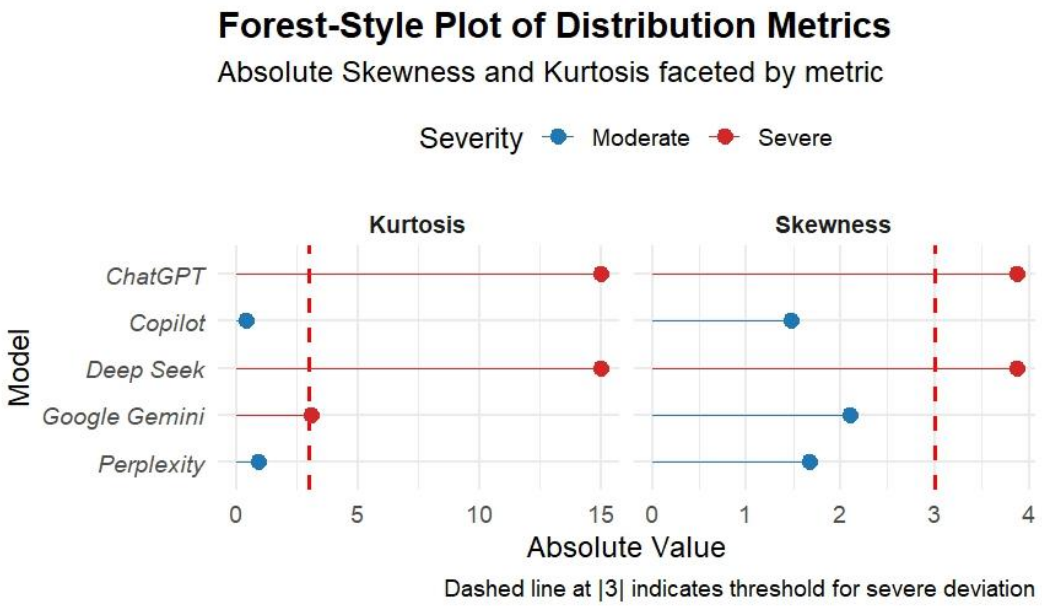
AI Model	Total Score (out of 90)	Percentage (%)	Mean Score	Median	Std. Deviation	Variance	Range	Skewness	Kurtosis
ChatGPT 3.5	88	97.77	5.87	6.00	0.516	0.267	02	−3.873	15.000
DeepSeek AI	86	95.55	5.73	6.00	1.033	1.067	04	−3.873	15.000
Perplexity AI	84	93.33	5.60	6.00	0.828	0.686	02	−1.672	0.897
Google Gemini	78	86.66	5.20	6.00	1.781	3.171	05	−2.098	3.107
Bing Copilot	73	81.11	4.86	6.00	2.066	4.267	05	−1.478	0.392

Figure 1. *Presents a Faceted Dot Plot Comparing the Total Performance Scores and Recommendation Accuracy of Five AI Models.*



The standard deviations in scores differed, with ChatGPT exhibiting the lowest variability (0.516), signifying highly consistent performance across the scenarios. In contrast, Copilot demonstrated the highest standard deviation (2.066), indicating greater variability in its performance across various neurological conditions. The persistent negative skewness values (from −1.478 for Copilot to −3.873 for ChatGPT and DeepSeek) across all models indicate that the score distribution was biased towards the higher end, suggesting that the AI platforms generally exhibited strong performance, with a greater concentration of scores near the maximum possible value rather than lower scores. The elevated kurtosis values for ChatGPT and DeepSeek (15.000) indicate a sharply peaked distribution with substantial tails, signifying a tendency for scores to cluster around the mean, accompanied by a limited number of extreme outliers (Figure 2).

Figure 2. *The Forest Plot Compares Absolute Skewness and Kurtosis Across AI Models (ChatGPT, Copilot, Etc.), With a Dashed Line at 3 Indicating Severe Deviation. It Highlights Distribution Irregularities, Aiding Model Reliability Assessment.*



Normality Testing

In order to find the right statistical tests to compare the performance of AI platforms, we used the Kolmogorov-Smirnov and Shapiro-Wilk tests to see if the distribution of the 'Marks' (scores) was normal for each AI platform.

Table 2 demonstrates that both the Kolmogorov-Smirnov and Shapiro-Wilk tests produced significant values (p < .001) for all AI platforms. A p-value below .05 signifies a statistically significant divergence from a normal distribution. Thus, these findings validate that the performance metrics for each AI model exhibited a nonnormal distribution. This discovery is significant as it requires the application of nonparametric statistical methods, such as the Kruskal-Wallis H test, for future comparisons among the AI systems, given that these tests do not presume a normal distribution of the data.

Characteristics of Verbal Output

In addition to accuracy, the study analyzed the characteristics of the verbal output produced by each AI platform. Table 3 presents the aggregate word count and the average word count per response, including their corresponding standard deviations.

DeepSeek AI regularly generated the most extensive responses, averaging 488.4 words each response. ChatGPT ranked as the second most verbose, averaging 429.60 words per response, while Google Gemini produced roughly 379.06 words for each response. Conversely, Copilot and Perplexity offered significantly more short responses, with average word counts of 238.86 and 240.33, respectively. The standard deviation of DeepSeek AI (48.295) signifies a very uniform response length, despite an elevated word count, indicating a methodical approach to thorough responses. Perplexity, albeit generating shorter replies, demonstrated marginally greater variability (59.513) in response length than Copilot. The discrepancies in vocal output may indicate differing degrees of complexity, explanatory depth, or verbosity in the AI's presentation of diagnostic and therapeutic information.

Therapeutic Recommendation Accuracy

Assessing the accuracy of each AI platform's therapeutic advice was a crucial component of the study. Table 5 demonstrates outstanding efficacy in therapeutic recommendations by DeepSeek AI, ChatGPT 3.5, and Google Gemini, each attaining a flawless 100% accuracy rate by delivering 15 accurate recommendations out of 15, without any erroneous suggestions. Perplexity AI exhibited robust performance, attaining a 93.3% accuracy rate

with 14 correct recommendations and merely 1 erroneous one. Copilot exhibited the lowest accuracy in this area, achieving 12 correct and 3 incorrect recommendations, culminating in an 80% accuracy rate (Figure 3). This finding underscores the exceptional reliability of leading AI models in producing precise treatment recommendations, an essential element in clinical neurology.

Table 2
Normality Tests (Kolmogorov–Smirnov and Shapiro–Wilk)

AI Model	KS Stat	KS df	KS Sig.	SW Stat	SW df	SW Sig.
ChatGPT 3.5	0.535	15	.000	0.284	15	.000
DeepSeek AI	0.535	15	.000	0.284	15	.000
Perplexity AI	0.485	15	.000	0.499	15	.000
Google Gemini	0.473	15	.000	0.503	15	.000
Bing Copilot	0.442	15	.000	0.573	15	.000

Table 3
Word Count Summary of AI Responses

AI Model	Total Word Count	Mean Words per Response	Standard Deviation
DeepSeek AI	7326	488.40	48.295
ChatGPT 3.5	6444	429.60	70.632
Google Gemini	5686	379.06	57.231
Perplexity AI	3605	240.33	59.513
Bing Copilot	3583	238.86	40.059

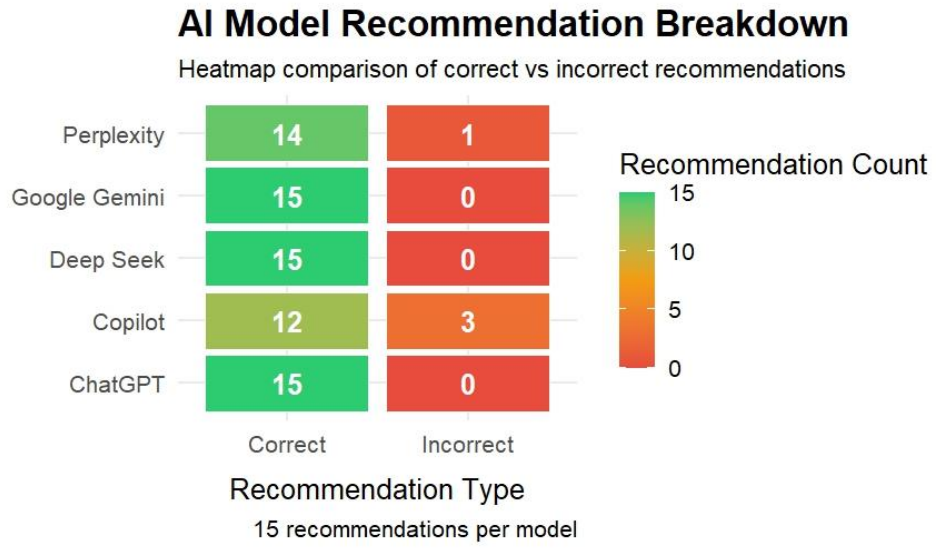
Table 4
Diagnostic Accuracy Overview

AI Model	Correct Answers (out of 45)	Average Score	Incorrect Recommendations
ChatGPT 3.5	44	2.93	01
DeepSeek AI	43	2.86	02
Perplexity AI	42	2.80	03
Google Gemini	38	2.53	07
Bing Copilot	32	2.13	13

Table 5
Therapeutic Recommendation Accuracy

AI Model	Correct Recommendations	Incorrect Recommendations	Accuracy (%)
ChatGPT 3.5	15	0	100%
DeepSeek AI	15	0	100%
Google Gemini	15	0	100%
Perplexity AI	14	1	93.3%
Bing Copilot	12	3	80%

Figure 3. *The Heatmap Shows AI Models' Recommendation Accuracy, With ChatGPT, Gemini, and DeepSeek Scoring Perfectly (15/15 Correct), While Copilot (12/15) and Perplexity (14/15) Had Minor Errors.*



Inferential Statistical Analysis

A Chi-square test of independence and a Kruskal-Wallis H test were performed to thoroughly evaluate the statistical significance or random chance of the observed performance differences across the AI systems.

A Chi-square test of independence was conducted to examine the relationship between the used AI platform and the achieved diagnostic accuracy. The study produced a nonsignificant result, $\chi^2(4, N = 75) = 4.245$, with a *p*-value of .374. Since the *p*-value (.374) is significantly above the traditional significance threshold of $\alpha = .05$, the null hypothesis of no significant relationship between the AI platform and diagnostic accuracy is confirmed. This finding indicates that, statistically, the distribution of diagnostic accuracy does not significantly differ among the various AI platforms assessed in this study.

Additionally, a nonparametric independent-samples Kruskal-Wallis H test was utilized to evaluate the overall performance scores (marks) among the various categories of AI platforms, as the normality tests (Table 2) revealed a nonnormal data distribution. The synopsis of this hypothesis test is defined in Table 6 and Table 7. There was no statistically significant difference in the AI platform performance scores, according to the Kruskal-Wallis H test results (Table 6 and 7), with a $\chi^2(4, N = 75) = 3.875$ and a *p*-value of .423. The *p*-value (.423) above the preestablished significance level of $\alpha = .05$, supporting the null hypothesis that the distribution of marks is consistent across all AI categories. This statistical result indicates that, although there are numerical changes in total scores and percentages (as illustrated in Table 1), these differences are insufficient to be deemed statistically significant. Consequently, the comprehensive statistical analysis indicates that the performance of the assessed AI platforms in aiding neurological diagnosis and suggestions, within the parameters of this study, did not exhibit significant differences (Figure 4). Therefore, in accordance with typical nonparametric analysis protocols, no post hoc pairwise Mann-Whitney U tests were performed due to the nonsignificant overall test outcome.

Table 6

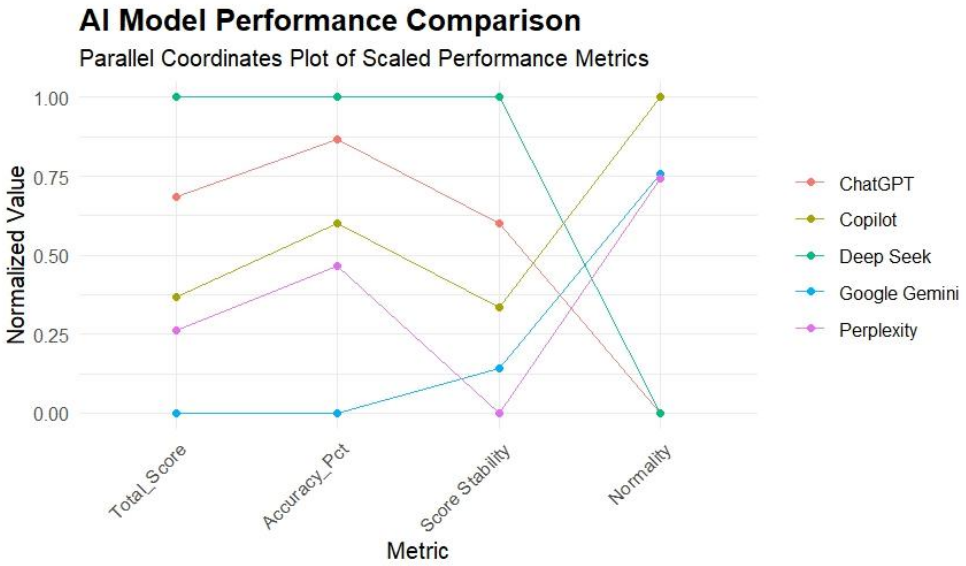
Kruskal-Wallis H Test Summary (Hypothesis Test)

Null Hypothesis	Test Used	*p*-value	Decision
The distribution of marks is the same across AI platforms	Kruskal-Wallis Test	.423	Retain null hypothesis

Table 7
Kruskal-Wallis Detailed Output

Statistic	Value
Total N	75
Test Statistic (χ^2)	3.875
Degrees of Freedom	4
Asymptotic Significance	0.423
Adjusted for ties	Yes
Post hoc comparisons	Not performed

Figure 4. *Parallel Coordinates Plot Comparing the Performance of Five AI Language Models (ChatGPT, Copilot, DeepSeek, Google Gemini, and Perplexity) Across Four Key Metrics: Total Score, Accuracy Percentage, Score Stability (Inverse of Standard Deviation), and Normality (Based on the Shapiro-Wilk Statistic).*



Note. All values are normalized to a 0–1 scale to enable direct comparison across dimensions. This visualization highlights the multidimensional variability in model performance.

Discussion

One of the most exciting developments in contemporary medicine is the incorporation of AI into clinical decision-making. Our side-by-side comparison of five well-liked AI tools—ChatGPT, Microsoft Copilot, DeepSeek, Google Gemini, and Perplexity—demonstrates considerable differences in their capacity for making precise neurological diagnoses, suggesting relevant diagnostic tests, and proposing treatment plans.

Performance Overview

Of the models that were tested, ChatGPT was the most precise with 97.77%, closely followed by DeepSeek with 95.55% and Perplexity with 93.33%. All these models consistently answered the whole of the prompt's three parts: diagnosis, next diagnostic test, and therapeutic or molecular basis. Their high performance indicates that they are adept at mimicking multifaceted clinical reasoning, especially if provided with well-formatted case vignettes from *Case Files: Neurology, Third Edition*.

The responses generated by Google Gemini (86.66%) and Copilot (81.11%) were more inconsistent. Although both models accurately diagnosed numerous instances, they often left out critical diagnostic or therapeutic details, particularly for less well-known conditions such as Creutzfeldt-Jakob disease (CJD), chronic inflammatory demyelinating polyneuropathy (CIDP), and amyotrophic lateral sclerosis (ALS).

The results of this study are in agreement with previous studies that although large language models (LLMs) may have junior resident-level performance in certain diagnostic tasks, their reliability is highly conditional on both the condition's complexity as well as the input question's specificity (Mota et al., 2023; Surianarayanan et al., 2023).

Comparative Strategies Throughout Models

In spite of the variation in accuracy, statistical analysis conducted using the Kruskal-Wallis test showed the absence of a significant difference overall among the AI platforms (p = .423), which meant all the models operated within a fairly comparable range (Sahu et al., 2022). Qualitative analysis of the structure and depth of the responses, however, showed noticeable differences:

Both ChatGPT and DeepSeek gave longer, more explanatory responses, which contributed to higher accuracy but might be daunting for users seeking concise summaries. Perplexity and Gemini favored conciseness, enhancing readability but sometimes at the cost of diagnostic accuracy. Copilot struggled to generate highly structured, evidence-based recommendations, hence exposing potential flaws in its store of medical information. The results emphasize the necessity of achieving a balance between precision and practicality in AI-supported diagnostic processes. Although extensive outputs may enhance accuracy, they must concurrently remain interpretable and actionable for healthcare professionals (Kaur et al., 2020).

Advantages of Artificial Intelligence in Neurological Diagnosis

This research determined the following benefits:

Speed and Efficiency. AI models processed each case in a matter of seconds, enabling quick hypothesis generation and differential narrowing—vital in time-critical neurological emergencies such as subarachnoid hemorrhage or Guillain-Barré syndrome.

Consistency. Unlike human experts who might be impacted by fatigue or cognitive bias, AI provides consistent responses based on available information.

Accessibility. These tools provide high-level information to nonexpert individuals and in resource-poor environments, with the potential to close gaps in global neurology care (Rudie et al., 2019).

Rare Disease Support. AI systems excelled at recognizing patterns for rare diseases, where experience might be sparse (Shen et al., 2019).

Limitations and Ethical Considerations

For all their potential, these AI models have certain limitations:

Lack of Clinical Contextualization. AI lacks in showing bedside manner, expressing empathy, and reading nonverbal cues that are crucial in performing neurological assessments. Excessive dependence on training data results in responses generated from large yet static datasets, which may lack more current diagnostic criteria or treatment protocols.

Regulatory and Liability Issues. The use of AI in clinical decision-making raises unclear questions about validation, oversight, and liability (Segato et al., 2020).

Also, our findings point to human-AI collaboration rather than complete automation. AI will be an additive, complementary tool, augmenting—rather than replacing—the expertise of experienced neurologists (Tyagi & Sharma, 2021).

Comparison With Current Literature

Our results align with prior research showing high diagnostic performance of LLMs in controlled settings (Kaur et al., 2020; Rudie et al., 2019). For instance, Shen et al. (2019) showed that AI models were superior to or as good as doctors across a range of diagnostic domains, including internal medicine and dermatology (World Health Organization, 2023). Similarly, Rudie et al. (2019) highlighted the growing contribution of AI to neuro-oncology, with particular reference to image interpretation and pattern recognition (Mota et al., 2023).

But whereas most of the previous studies concentrated on image-based diagnosis or symptom checkers, our research tested the capacity of AI to integrate multidomain clinical reasoning—a key to

neurology, where syndromic cognition is likely to precede a particular diagnosis.

Future Trajectories

To maximize the potential of AI for neurology, future efforts should concentrate on domain-specific fine-tuning, which involves training models with meticulously chosen medical information and case studies specific to neurological disorders.

- Explainability and Transparency: developing methods to monitor how AI arrives at a diagnosis, engendering trust and aiding clinicians in verifying results.
- Incorporation into Clinical Practices: integration of AI tools into electronic health records (EHRs) and decision support systems allows for their seamless use during patient encounters.
- Human-in-the-loop Systems: creating collaborative interfaces that combine AI efficiency with clinician experience to improve outcomes (Alyami et al., 2024; Rahman et al., 2024).

Conclusion

This comparative evaluation demonstrates that publicly accessible AI platforms, specifically DeepSeek (95.6%) and ChatGPT (97.8%), exhibit high diagnostic accuracy and therapeutic recommendation efficacy in standardized neurological cases, comparable to junior residents. Though there were no statistically significant differences between the platforms (p > .05), there were noticeable qualitative differences. DeepSeek and ChatGPT offered more in-depth explanations in their responses, whereas Perplexity and Copilot prioritized being brief, sometimes at the expense of being completely thorough.

AI models demonstrated exceptional performance in the areas of speed, consistency, and rare-disease recognition. However, they are reliant on immutable training data and lack clinical contextualization, such as empathy and nonverbal assessment. Thus, their function should be supplementary, enhancing rather than replacing the expertise of the clinician.

Author Disclosure

AI tools were used solely for language editing and stylistic refinement of the manuscript. No AI tools were used for study design, data collection, data analysis, or interpretation of results. The authors take full responsibility for the accuracy, integrity, originality, and ethical compliance of the entire manuscript.

References

AbuAlrob, M. A., & Mesraoua, B. (2024). Harnessing artificial intelligence for the diagnosis and treatment of neurological emergencies: A comprehensive review of recent advances and future directions. *Frontiers in Neurology, 15*, Article 1485799. https://doi.org/10.3389/fneur.2024.1485799

Alhejaily, A.-M. G. (2025). Artificial intelligence in healthcare. *Biomedical Reports, 22*(1), Article 11. https://doi.org/10.3892/br.2024.1889

Alyami, M. S. M., Alyami, M. M. M., Al Khuraim, H. A. M., Alsalem, A. M. S., Alrayshan, H. A. M., Albakri, K. A. M., Alsaqran, Q. N., Alyami, H. S., Alzamanan, A. S., Alharbi, F. M., & Alharbi, F. M. (2024). Integrating artificial intelligence across medical clinics: strengthening collaborative efforts for improved patient outcomes. *Journal of Ecohumanism, 3*(7), 2691–2698. https://doi.org/10.62754/joe.v3i7.4668

Dipankar, P., Salazar, D., Dennard, E., Mohiyuddin, S., & Nguyen, Q. C. (2025). Artificial intelligence based advancements in nanomedicine for brain disorder management: An updated narrative review. *Frontiers in Medicine, 12*, Article 1599340. https://doi.org/10.3389/fmed.2025.1599340

Kalani, M., & Anjankar, A. (2024). Revolutionizing neurology: The role of artificial intelligence in advancing diagnosis and treatment. *Cureus, 16*(6), Article e61706. https://doi.org/10.7759/cureus.61706

Kandel, A. (2025). Addressing neurological inequities in developing countries: Challenges and strategic solutions. *Sarvodaya International Journal of Medicine, 1*(1), 1–11. https://doi.org/10.4103/SIJM.SIJM_2_24

Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., El-Sappagh, S., Islam, M. S., & Islam, S. M. R. (2020). Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives. *IEEE Access, 8*, 228049–228069. https://doi.org/10.1109/ACCESS.2020.3042273

Li, X., Zhang, L., Yang, J., & Teng, F. (2024). Role of artificial intelligence in medical image analysis: A review of current trends and future directions. *Journal of Medical and Biological Engineering, 44*(2), 231–243. https://doi.org/10.1007/s40846-024-00863-x

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon, 10*(4), Article e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

Mota, A. L., Ferraciolli, S. F., Ayres, A. S., Polsin, L. L. M., da Costa Leite, C., & Kitamura, F. (2023). AI and big data for intelligent health: Promise and potential. In H. Sakly, K. Yeom, S. Halabi, M. Said, J. Seekins, & M. Tagina (Eds.), *Trends of artificial intelligence and big data for e-health* (pp. 1–14). Springer. https://doi.org/10.1007/978-3-031-11199-0_1

Nguyen, T. V., & Vo, N. (2024). *Using traditional design methods to enhance AI-driven decision making*: IGI Global. https://doi.org/10.4018/979-8-3693-0639-0

Onciul, R., Tataru, C.-I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M. P., & Toader, C. (2025). Artificial intelligence and neuroscience: Transformative synergies in brain research and clinical applications. *Journal of Clinical Medicine, 14*(2), 550. https://doi.org/10.3390/jcm14020550

Oyeniyi, J., & Oluwaseyi, P. (2024). Emerging trends in AI-powered medical imaging: Enhancing diagnostic accuracy and treatment decisions. *International Journal of Enhanced Research in Science, Technology & Engineering, 13*(4), 81–94. https://doi.org/10.55948/IJERSTE.2024.0412

Rahman, M. H., Hossan, K. M. R., Uddin, M. K. S., & Hossain, M. D. (2024). Improving collaborative interactions between humans and artificial intelligence to achieve optimal patient

outcomes in the healthcare industry. *SSRN,* Article 5029975. https://doi.org/10.2139/ssrn.5029975

Rashid, M., & Sharma, M. (2025). AI-assisted diagnosis and treatment planning—A discussion of how AI can assist healthcare professionals in making more accurate diagnoses and treatment plans for diseases. In R. Singh, A. Gehlot, N. Rathour, & S. V. Akram (Eds.), *AI in disease detection: Advancements and applications* (pp. 313–336). Wiley-IEEE Press. https://doi.org/10.1002/9781394278695.ch14

Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C., & Mohan, S. (2019). Emerging applications of artificial intelligence in neuro-oncology. *Radiology, 290*(3), 607–618. https://doi.org/10.1148/radiol.2018181928

Sahu, M., Gupta, R., Ambasta, R. K., & Kumar, P. (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Progress in Molecular Biology and Translational Science, 190*(1), 57–100. https://doi.org/10.1016/bs.pmbts.2022.03.002

Salammagari, A. R. R., & Srivastava, G. (2024). Artificial intelligence in healthcare: Revolutionizing disease diagnosis and treatment planning. *International Journal of Research in Computer Applications and Information Technology, 7*(1), 41–53.

Segato, A., Marzullo, A., Calimeri, F., & De Momi, E. (2020). Artificial intelligence for brain diseases: A systematic review. *APL Bioengineering, 4*(4), Article 041503. https://doi.org/10.1063/5.0011697

Shen, J., Zhang, C. J., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Medical Informatics, 7*(3), Article e10010. https://doi.org/10.2196/10010

Shokran, M., Islam, M. S., & Ferdousi, J. (2025). Harnessing AI adoption in the workforce a pathway to sustainable competitive advantage through intelligent decision-making and skill transformation. *American Journal of Economics and Business Management, 8*(3), 954–976. https://globalresearchnetwork.us/index.php/ajebm/article/view/3355

Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E., & Hewage, C. (2023). Convergence of artificial intelligence and neuroscience towards the diagnosis of neurological disorders—A scoping review. *Sensors, 23*(6), 3062. https://doi.org/10.3390/s23063062

Toy, E. C., Simpson, E. P., Mancias, P., Furr Stimming, E. E. (2018). *Case Files: Neurology, Third Edition*. McGraw Hill.

Tyagi, Y., & Sharma, P. K. (2021). Artificial intelligence: An emerging approach in healthcare. In *Artificial intelligence* (pp. 71–92). CRC Press. https://doi.org/10.3389/fdgth.2025.1644041

Valerio, J. E., Aguirre Vera, G. d. J., Fernandez Gomez, M. P., Zumaeta, J., & Alvarez-Pinzon, A. M. (2025). AI-driven advances in Parkinson's disease neurosurgery: Enhancing patient selection, trial efficiency, and therapeutic outcomes. *Brain Sciences, 15*(5), 494. https://doi.org/10.3390/brainsci15050494

World Health Organization. (2023). *WHO framework for meaningful engagement of people living with noncommunicable diseases, and mental health and neurological conditions*. World Health Organization.

Yang, R., Liu, X., Zhao, Z., Zhao, Y., & Jin, X. (2025). Burden of neurological diseases in Asia, from 1990 to 2021 and its predicted level to 2045: A global burden of disease study. *BMC Public Health, 25*(1), Article 706. https://doi.org/10.1186/s12889-025-21928-9

Zeb, S., Nizamullah, F., Abbasi, N., & Fahad, M. (2024). AI in healthcare: Revolutionizing diagnosis and therapy. *International Journal of Multidisciplinary Sciences and Arts, 3*(3), 118–128. https://doi.org/10.47709/ijmdsa.v3i3.4546

Received: August 3, 2025
Accepted: September 23, 2025
Published: March 31, 2026