



Article

Specialized Large Language Model Outperforms Neurologists at Complex Diagnosis in Blinded Case-Based Evaluation

Sami Barrit ^{1,2,3,4,*}, Nathan Torcida ^{4,5,†}, Aurelien Mazeraud ^{6,7}, Sebastien Boulogne ⁸, Jeanne Benoit ⁹, Timothée Carette ¹⁰, Thibault Carron ¹¹, Bertil Delsaut ^{5,12}, Eva Diab ¹³, Hugo Kermorvant ¹⁴, Adil Maarouf ^{15,16}, Sofia Maldonado Sloodjes ^{17,18}, Sylvain Redon ¹⁹, Alexis Robin ²⁰, Sofiene Hadidane ²¹, Vincent Harlay ²², Vito Tota ²³, Tanguy Madec ²⁴, Alexandre Niset ^{4,25,26}, Mejdeddine Al Barajraji ^{4,27}, Joseph R. Madsen ³, Salim El Hadwe ^{1,4,28}, Nicolas Massager ^{1,2}, Stanislas Lagarde ^{29,30,‡} and Romain Carron ^{4,29,31,‡}

- ¹ Neurosurgery, Université Libre de Bruxelles, 1070 Brussels, Belgium; se471@cam.ac.uk (S.E.H.); nmessage@chu-tivoli.be (N.M.)
- ² Neurosurgery, CHU Tivoli, 7110 La Louvière, Belgium
- ³ Neurodynamics Laboratory, Department of Neurosurgery, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA; joseph.madsen@childrens.harvard.edu
- ⁴ Science, New York, NY 10013, USA; alexandre.niset@gmail.com (A.N.); mejdi.albarajraji@gmail.com (M.A.B.); romain.carron@ap-hm.fr (R.C.)
- ⁵ Neurology, Université Libre de Bruxelles, 1050 Brussels, Belgium; bertil.delsaut@gmail.com
- ⁶ Anesthésie-Réanimation, GHU Paris, Pôle Neuro, 75014 Paris, France; aurelien.mazeraud@gmail.com
- ⁷ Neurosciences, Université de Paris, 75006 Paris, France
- ⁸ Neurophysiology and Epileptology, Université de Lyon, 69007 Lyon, France; sebastien.boulogne@chu-lyon.fr
- ⁹ Neurology, CHU de Nice, Université Côte d'Azur, UMR2CA, 06000 Nice, France; benoit.j@chu-nice.fr
- ¹⁰ Neurology, Université Catholique de Louvain, Clinique Saint-Pierre Ottignies, 1348 Louvain-la-Neuve, Belgium; timothee.carette@gmail.com
- ¹¹ LIP6, CNRS, Sorbonne Université, 75005 Paris, France; thibault.carron@gmail.com
- ¹² Neurology, CHU Tivoli, 7110 La Louvière, Belgium
- ¹³ Clinical Neurophysiology, CHU Amiens Picardie, CHIMERE UR 7516 UPJV, 80054 Amiens, France; diab.eva@chu-amiens.fr
- ¹⁴ Neurophy Lab, Université Libre de Bruxelles, 1050 Brussels, Belgium; hugokermorvant@gmail.com
- ¹⁵ Neurology, La Timone Hospital, AP-HM, 13385 Marseille, France; adil.maarouf@ap-hm.fr
- ¹⁶ Department of Neurology, Maladie Inflammatoire du Cerveau et de la Moelle Epinière (MICeME), Aix Marseille Université (AMU), CNRS, CRMBM, 13385 Marseille, France
- ¹⁷ Department of Neurology, Universitair Ziekenhuis Brussel (UZ Brussel), 1090 Brussels, Belgium
- ¹⁸ NEUR Research Group, Vrije Universiteit Brussel (VUB), 1090 Brussels, Belgium
- ¹⁹ Evaluation and Treatment of Pain, FHU INOVRAIN, La Timone Hospital, AP-HM, 13385 Marseille, France; sylvain.redon@ap-hm.fr
- ²⁰ Neurology, CHU Grenoble, 38700 Grenoble, France
- ²¹ Cabinets de Neurologie d'Allauch et Plan de Cuques, 13190 Allauch, France; sofien.hadidane@gmail.com
- ²² Neuro-Oncology, AMU, La Timone Hospital, AP-HM, 13005 Marseille, France; vincent.harlay@ap-hm.fr
- ²³ Neurology, CHU Helora, 7000 Mons, Belgium
- ²⁴ Neurology, Hospital of Noumea, 98800 Nouméa, France; tanguy.madec.raybaud@gmail.com
- ²⁵ Emergency Medicine, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium
- ²⁶ Pediatric Intensive Care Unit, Cliniques Universitaires Saint-Luc, 1200 Brussels, Belgium
- ²⁷ Département des Neurosciences Cliniques, Centre Hospitalier Universitaire Vaudois (CHUV), 1005 Lausanne, Switzerland
- ²⁸ Clinical Neuroscience, University of Cambridge, Cambridge CB2 1TN, UK
- ²⁹ AMU, INSERM, Institut Neurosciences des Systèmes (INS), 13005 Marseille, France; stanislas.lagarde@gmail.com
- ³⁰ APHM, Timone Hospital, Epileptology and Cerebral Rhythmology, 13005 Marseille, France
- ³¹ Stereotactic and Functional Neurosurgery, La Timone Hospital, AP-HM, 13385 Marseille, France
- * Correspondence: samibarrit@gmail.com
- † These authors contributed equally to this work.
- ‡ These authors also contributed equally to this work.



Academic Editor: Julian Keenan

Received: 5 March 2025

Revised: 24 March 2025

Accepted: 25 March 2025

Published: 27 March 2025

Citation: Barrit, S.; Torcida, N.; Mazeraud, A.; Boulogne, S.; Benoit, J.; Carette, T.; Carron, T.; Delsaut, B.; Diab, E.; Kermorvant, H.; et al. Specialized Large Language Model Outperforms Neurologists at Complex Diagnosis in Blinded Case-Based Evaluation. *Brain Sci.* **2025**, *15*, 347. <https://doi.org/10.3390/brainsci15040347>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Background/Objectives: Artificial intelligence (AI), particularly large language models (LLMs), has demonstrated versatility in various applications but faces challenges in specialized domains like neurology. This study evaluates a specialized LLM's capability

and trustworthiness in complex neurological diagnosis, comparing its performance to neurologists in simulated clinical settings. **Methods:** We deployed GPT-4 Turbo (OpenAI, San Francisco, CA, US) through Neura (Sciense, New York, NY, US), an AI infrastructure with a dual-database architecture integrating “long-term memory” and “short-term memory” components on a curated neurological corpus. Five representative clinical scenarios were presented to 13 neurologists and the AI system. Participants formulated differential diagnoses based on initial presentations, followed by definitive diagnoses after receiving conclusive clinical information. Two senior academic neurologists blindly evaluated all responses, while an independent investigator assessed the verifiability of AI-generated information. **Results:** AI achieved a significantly higher normalized score (86.17%) compared to neurologists (55.11%, $p < 0.001$). For differential diagnosis questions, AI scored 85% versus 46.15% for neurologists, and for final diagnosis, 88.24% versus 70.93%. AI obtained 15 maximum scores in its 20 evaluations and responded in under 30 s compared to neurologists’ average of 9 min. All AI-provided references were classified as relevant with no hallucinatory content detected. **Conclusions:** A specialized LLM demonstrated superior diagnostic performance compared to practicing neurologists across complex clinical challenges. This indicates that appropriately harnessed LLMs with curated knowledge bases can achieve domain-specific relevance in complex clinical disciplines, suggesting potential for AI as a time-efficient asset in clinical practice.

Keywords: artificial intelligence; large language models; neurological diagnosis; clinical decision support

1. Introduction

Artificial intelligence (AI) has become an instrumental force across multiple sectors, notably in healthcare [1] and biomedical research [2]. Within this expansive realm, large language models (LLMs) have garnered attention for their proficiencies in natural language processing (NLP). These models have demonstrated versatility in diverse, broad applications, most recently exemplified by the prominent advent of conversational agents [3–5]. However, their deployment in specialized scientific domains, particularly medicine, is distinctly challenging [6] due to the stringent constraints inherent to medical applications and the nuanced, discipline-specific considerations such domains entail [7]. Neurology—with its intricate clinical manifestations, neural substrates, and interdisciplinary integration—is a prime example of a complex and rapidly evolving expanse of knowledge that may be substantially embedded—and effectively encoded—in natural language. In practice, neurologists skillfully elicit detailed physical examination findings, which they then integrate with data from diverse diagnostic modalities through sophisticated clinical reasoning pathways defining high-stakes medical management.

Hence, a salient challenge resides in fine-tuning LLMs to achieve domain-specific relevance. Traditional fine-tuning methods are resource-intensive, requiring substantial computational and human capital [8]. Consequently, these methods are often feasible only for large-scale projects with considerable resources [9]. Another limitation of conventional LLM implementation is interpretability and transparency in information processing [10]—a critical requirement for verifiable information generation for medical and research purposes. Furthermore, while modern LLMs have expanded context windows, they still face attention degradation across long contexts, potentially limiting their effectiveness in complex, data-rich environments typical of healthcare and research [11,12]. Here, we evaluate a specialized

LLM's capability and trustworthiness in complex neurological diagnosis, comparing its performance to neurologists in simulated clinical settings.

2. Materials and Methods

2.1. AI System

Neura (Science, New York, NY, USA) is a solution deploying an LLM with custom parameters and prompt engineering on curated corpora with extended contexts for advanced grounding through retrieval-augmented generation (RAG) [13]. This solution is predicated on a dual-database architecture integrating both 'long-term memory' (LTM) and 'short-term memory' (STM) components. The LTM serves as the repository for domain-specific knowledge. It employs an agnostic, vectorized approach enabled by text embeddings generated from parsed source texts [14]. The STM captures the setting and conversational history between the user and the LLM, thereby adding a layer of contextual knowledge. The STM is implemented using a non-relational database [15], ensuring real-time accessibility and state persistence of conversational data. Information retrieval is optimized in speed and accuracy through a single-stage filtering process, integrating vector and metadata indexes into a unified structure [16] that integrates vector and metadata indexes into a unified structure, enabling simultaneous semantic similarity matching and exact term identification. This approach reduces computational overhead while maintaining context sensitivity. This dual-database architecture with LTM vectorization and STM context capture aims to address attention degradation challenges in long clinical contexts. Source tracking is enabled, culminating in actionable, standardized references for the end-user and ensuring verifiability of answer accuracy. For this study, we deployed a state-of-the-art LLM, GPT-4 Turbo (OpenAI, San Francisco, CA, USA) [5], on a prototype corpus curated for clinical neurology sourced from five comprehensive neurology textbooks [17–21], the neurologic disorders section of Merck's Manual (copyrighted) [22], and Wikipedia (open-source) [23] (Figure 1).

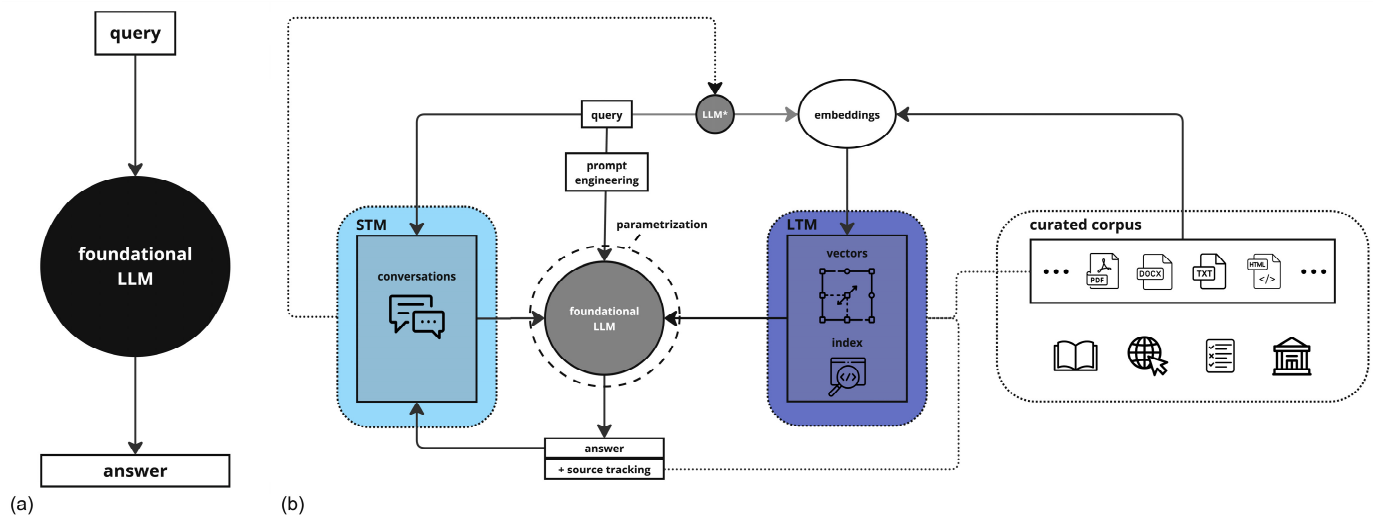


Figure 1. (a) Black-box use of LLM. (b) AI solution's architecture. LLM: large language model; * fine-tuned LLM for embedding generation; LTM: long-term memory; STM: short-term memory.

2.2. Diagnostic Challenges

Five representative clinical scenarios were adapted from peer-reviewed complex cases [24–28] to mirror the clinical practice through a two-tiered diagnostic approach. These cases were selected to represent diverse neurological subspecialties and encompass a spectrum of diagnostic complexity requiring the integration of clinical and paraclinical findings

(Table 1). The first tier required formulating and justifying an exhaustive differential diagnosis based on initial clinical presentation and findings. In the second tier, conclusive clinical information was provided to establish a definitive diagnosis (Table 2). We recruited senior residents and board-certified neurologists from teaching hospitals. Neurologists engaged in complex clinical reasoning to solve these diagnostic challenges, solely relying on intrinsic knowledge in the first tier—external resources were permitted in the second tier. All challenges were conducted via videoconferencing sessions, supervised by two investigators who provided documents presenting the cases (Supplementary S1), initial instructions, and procedural assistance. Answers with timing were recorded in text documents, which were subsequently collected and anonymized. AI undertook the challenges based on the same documents provided to neurologists. Answers were time-stamped and anonymized. Two senior academic neurologists, each responsible for residency training and educational programs at their respective universities, independently evaluated the answers, blinded to the involvement of AI as a participant. They employed a standardized scoring sheet (Supplementary S2), assigning points for precise and justified diagnoses and allowing bonus points for unexpected, relevant findings (Figure 2).

Table 1. Concise summaries of clinical presentations, neurologic fields, and final diagnoses for illustrative neurology cases demonstrating diagnostic challenges across diverse neurological disorders.

	Case Summary	Neurological Field	Final Diagnosis
Case 1	An 84-year-old Chinese woman with recurrent focal deficits, cognitive decline, multifocal cerebral artery constriction, and bilateral infarctions.	Neurovascular/neuro-oncology	Intravascular lymphoma
Case 2	A 44-year-old woman with hypothyroidism presenting with cognitive decline, headaches, confusion, and progressive multifocal white matter lesions.	CNS inflammation diseases	Susac disease
Case 3	A 50-year-old woman with systemic lupus erythematosus presenting with acute progressive left-sided weakness and sensory deficits, with imaging showing rapidly worsening right-hemispheric lesions.	Demyelinating diseases	Neuromyelitis optica spectrum disorder (NMOSD)
Case 4	A 65-year-old man with diabetes presenting with progressive asymmetric weakness, sensory deficits, areflexia, and weight loss.	PNS inflammation diseases	Neurosarcoidosis
Case 5	A 55-year-old man with extensive psychiatric history presenting with subacute-on-chronic cognitive decline, dysphagia, involuntary hyperkinetic movements, gait instability, and chronic transaminitis.	Movement disorders	Neuroacanthocytosis

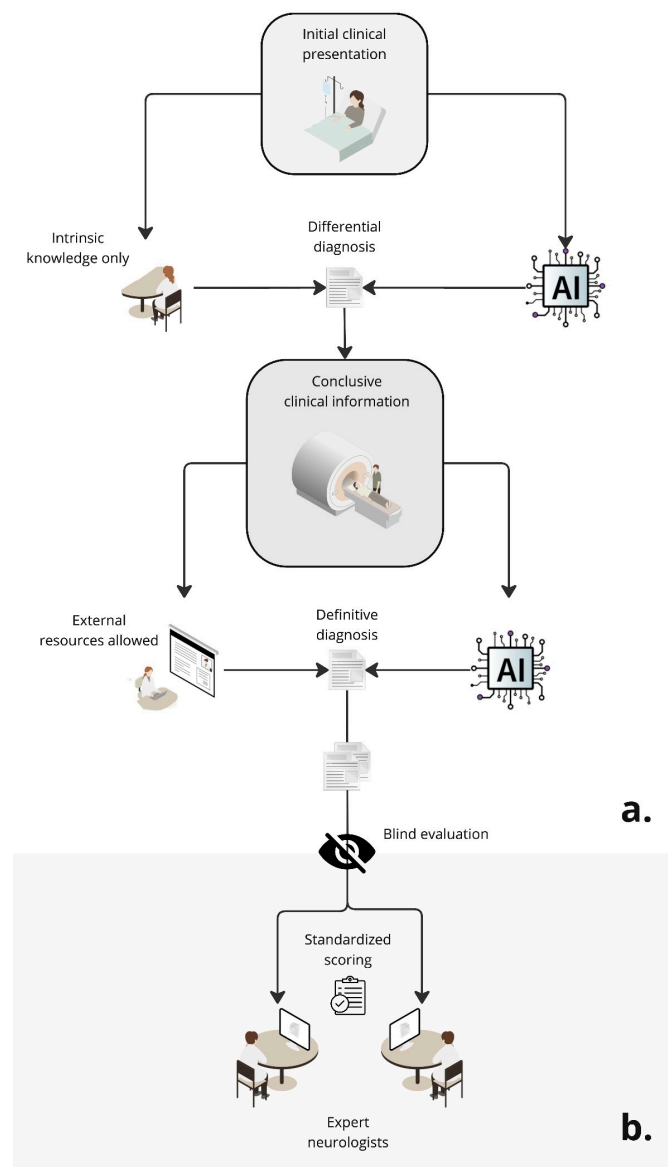


Figure 2. Overview of the simulated clinical diagnostic challenge: a videoconference setup for clinical scenarios simulations with human neurologists and specialized AI through grounding, parametrization, and prompt engineering; (a) shows the two-tiered diagnostic workflow from initial clinical presentation to differential diagnosis (using intrinsic knowledge only for human neurologists), followed by conclusive clinical information leading to definitive diagnosis (with external resources allowed for human neurologists); (b) represents the blind evaluation process where expert neurologists performed standardized scoring of anonymized responses from both human participants and AI.

Incorrect or risky conclusions incurred deductions, with a two-point loss yielding a null question score. Null scores from both evaluators constituted question failure. If multiple participants achieved the maximum score for a given question, the evaluator chose a preferred answer; conversely, if a single answer attained the maximum score, it was then defined as the highest score. In parallel, an independent investigator assessed the verifiability and reliability of the AI-generated information. This was achieved by classifying the references provided within the answers as relevant, irrelevant, or hallucinatory (i.e., incorrect or nonexistent) (Supplementary S3).

Table 2. Illustrative example of diagnostic workflow and scoring (Case 5). Structured clinical scenario highlighting the diagnostic reasoning process, scoring criteria for differential diagnoses (tier 1) and the final diagnosis (tier 2), demonstrating the study methodology.

Initial Presentation	Differential Diagnosis (Tier 1)	Ancillary Exams and Results	Final Diagnosis (Tier 2)
<ul style="list-style-type: none"> - Hyperkinetic involuntary movements affecting trunk and limbs, inability to suppress movements; - Chronic psychiatric conditions (PTSD, schizophrenia, anxiety, depression); - Chronic dysphagia, regurgitation, elevated transaminases (AST: 304 U/L, ALT: 154 U/L); - CK elevation (4753 IU/L, baseline ~600–800 IU/L); - MRI: bilateral caudate atrophy. 	<p>Evaluators awarded points (max 1 each) for justified differential diagnoses:</p> <ul style="list-style-type: none"> - Basal ganglia structural disorders (vascular, demyelinating); - Toxic-metabolic disorders (electrolyte derangements, hyperglycemia); - Drug-induced (antipsychotics, tardive dyskinesia); - Systemic autoimmune (lupus, APLS); - Hereditary (Huntington's, neuroacanthocytosis, Wilson); - Paraneoplastic (anti-CRMP5/NMDA). <p>Penalties for aberrant/harmful hypotheses (max – 2). Bonuses for plausible, justified alternative diagnoses (max + 2).</p>	<ul style="list-style-type: none"> - Serum copper and ceruloplasmin normal; - Extensive gastroenterologic and rheumatologic evaluations unrevealing; - MRI brain: bilateral caudate atrophy. 	<p>Neuroacanthocytosis (3 points awarded). (Alternative, less accurate diagnosis, Huntington's: 1 point.)</p> <p>Penalties possible for aberrant/harmful conclusions.</p>

2.3. Statistical Analysis

Descriptive statistics were calculated for the scores and times. For normalization, the maximum possible combined score for each question was determined by summing the highest score assigned by each evaluator. For any participant, we calculated the combined score from both evaluators for each question and then divided this by its maximum possible combined score. These resulting normalized scores were expressed as percentages, indicating the proportion of the maximum possible points each participant collected on a given question. We used the intraclass correlation coefficient (ICC) to measure consistency agreement for inter-rater reliability between evaluators. We compared the performance of the AI with that of neurologists using a linear mixed-effects model. Before analysis, we used residual plots, QQ plots, and Shapiro–Wilk tests to assess the assumptions of normality, homoscedasticity, and random effect structure. This model utilized average scores derived from the two evaluators as the dependent variable. Participant type (AI vs. human) was treated as a fixed effect, while variability across questions was modeled as random. The significance of the fixed effect was corroborated using an ANOVA with Satterthwaite's method for approximating degrees of freedom. We employed a Monte Carlo simulation (MCS) of 10,000 iterations to estimate the probabilities for AI achieving observed thresholds of maximum scores, highest scores, and preferred answers among its 20 scores by chance—assuming a uniform distribution of scores within each question's specific range across all participants. We set our alpha level threshold at 0.05 to determine statistical significance using two-tailed tests. All computations and visualizations were performed using R version 4.1.3, with the packages 'afex', 'eulerr', 'ggplot2', 'irr', 'lme4', and 'lmertest'.

3. Results

Of the 13 neurologists, 8 were board-certified. Challenges were conducted between March and October 2023. ICC(C,2) was found to be significant at 0.767 (95% CI [0.675, 0.833], $F(139,139) = 4.3$, p -value < 0.001). The residuals did not significantly deviate from normality ($W = 0.99327$, $p = 0.753$, Shapiro–Wilk test) as observed on the QQ plot, and plots of residuals versus fitted values supported homoscedasticity (Supplementary Figures S1 and S2). Additionally, random effects for participants and questions showed substantial variance (0.3789 and 0.3587, respectively, with a residual variance of 1.0584). Across all questions, AI achieved a significantly higher normalized score of 86.17% versus 55.11% for neurologists (SD = 14.81, range = 30.85–80.85; averages of 66.38% for residents and 48.07% for board-certified physicians—estimate = 1.46, Std. Error = 0.39, $df = 129$, $t = 3.75$, $p < 0.001$, linear mixed-effects model, and $F(1,129) = 14.021$, $p < 0.001$, type III ANOVA). For differential diagnosis questions, AI achieved a normalized score of 85% versus 46.15% for neurologists (SD = 15.24, range = 26.67–78.33; averages of 58.45% for residents and 39.40% for board-certified physicians). For final diagnosis, AI achieved a normalized score of 88.24% versus 70.93% for neurologists (SD = 17.36, range = 35.29–97.06; averages of 80.5% for residents and 64.87% for board-certified physicians) (Figure 3). AI performance notably decreased for differential diagnosis in Q2.1 and final diagnosis in Q4.2. In Q2.1, AI proposed broad diagnoses, including neoplastic and leukodystrophic conditions that evaluators deemed aberrant. In Q4.2, AI initially proposed neurosarcoidosis in the differential diagnosis but ultimately favored paraneoplastic neuropathy, a choice similarly made by most neurologists; this misalignment resulted in a null score due to the conservative scoring approach of one evaluator.

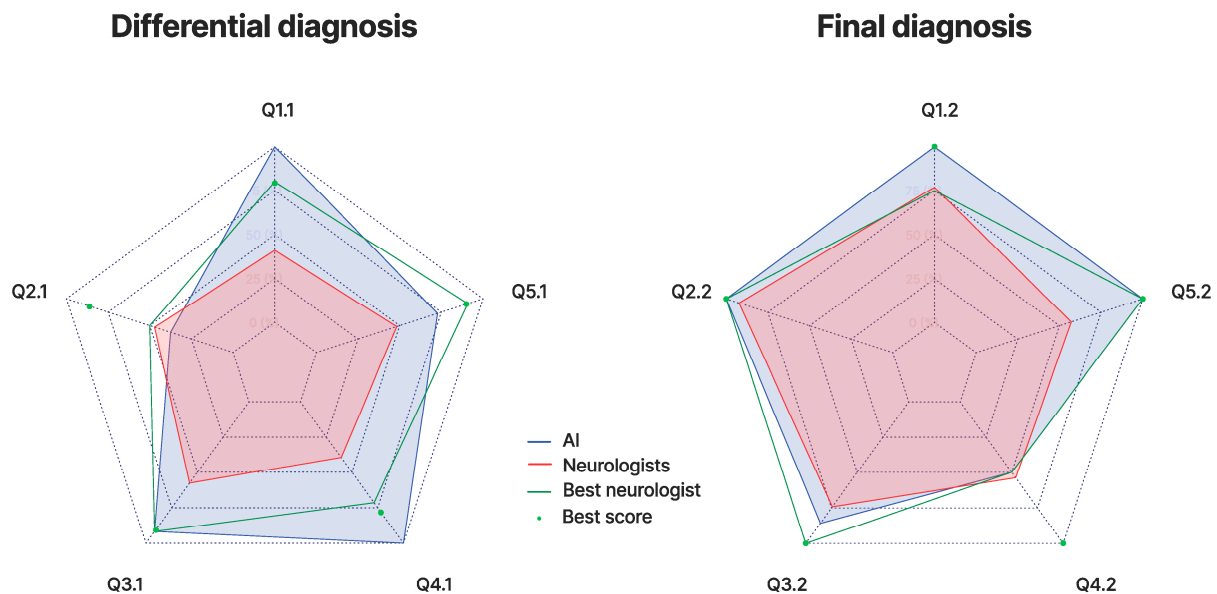


Figure 3. Radar charts of the performance across the differential diagnosis and final diagnosis challenges for AI, all neurologists collectively ('Neurologists'), the best-performing individual neurologist ('Best neurologist'), and the best individual score from all neurologists for each question ('Best score'). Each axis corresponds to a specific question, designated as Qx.y (where 'x' is the case number and 'y' is the tier), with scores normalized and depicted in increments of 25%.

The mean number of null scores and question failures was 2 and 0 for AI and 2 and 0.46 for neurologists (1 and 0.2 for residents and 2.625 and 0.625 for board-certified physicians). AI obtained 15 maximum scores (p -value < 0.001, MCS) in its 20 evaluations, 6 of the 8 highest scores (p -value < 0.001, MCS), and 4 of the 11 preferred answers from both evaluators (p -value = 0.03, MCS) (Figure 4).

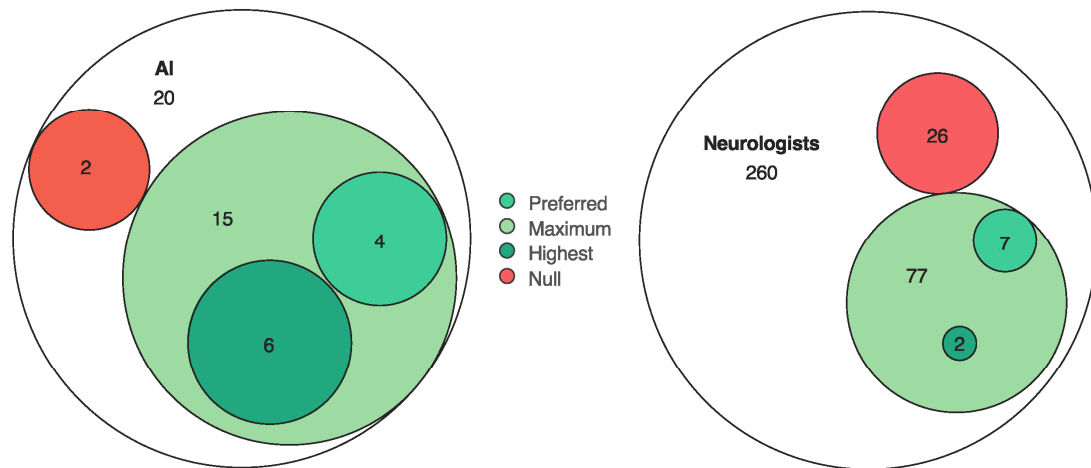


Figure 4. Euler diagrams representing AI and neurologists' answer attributes for AI, showing total and respective counts and proportions of evaluations in each category—null scores, maximum scores, highest scores, and preferred answers.

In comparison, the best neurologist, a resident, obtained a normalized score of 80.85%, with nine maximum scores, including two highest scores from one evaluator without a preferred answer and one null score. Neurologists' mean response times for differential and final diagnosis were 9.62 (SD = 4.47, range = 4–32) and 8.85 min (SD = 5.53, range = 1–30), compared to AI's mean times of 28.8 and 19.2 s, respectively (Figure 5).

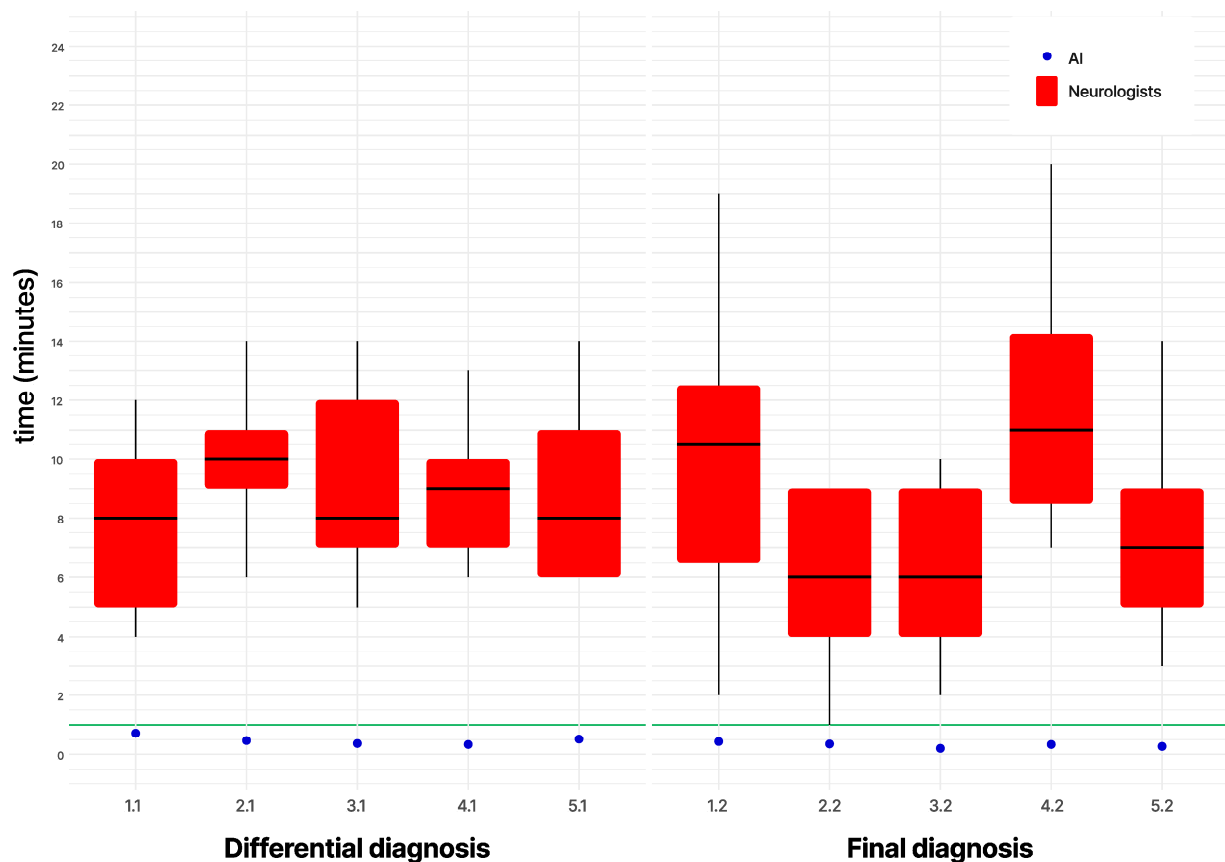


Figure 5. Box plots displaying the distribution of neurologists' response times, with AI's times as distinct points—horizontal green lines mark a one-minute reference.

All references provided by the AI were classified as relevant, and the generated information was deemed accurately derived from the cited sources. Despite the diverse

corpus including Wikipedia and Merck's Manual, source tracking revealed the AI system exclusively retrieved information from the peer-reviewed neurology textbooks for all diagnostic challenges. No instances of hallucinatory content were detected.

4. Discussion

In this blinded, controlled, comparative study, AI demonstrated superior diagnostic performance compared with a cohort of 13 neurologists across five complex clinical challenges, as evaluated by two academic neurologists. It achieved a significantly higher normalized score of 86.17% against the neurologists' 55.11%, with 15 maximum scores (including 6 highest scores) out of 20 evaluations. Across various levels of experience (residents and board-certified physicians) and types of diagnosis (differential and final), the AI consistently outperformed the neurologists. Interestingly, senior residents outperformed board-certified neurologists, a trend that might be attributed to the broader, ongoing training of the former, contrasting with the deep but narrow specialization of the latter. As this may reflect a judicious case selection for representing general neurology, it also raises the question of artificial and human acumen in niche or emerging domains of expertise.

In the diagnostic challenges, each tier encapsulated distinct aspects of clinical practice: the first tier encompassed bedside diagnosis, utilizing intrinsic knowledge and initial clinical presentation, while the second tier, extending to broader clinical investigation, allowed the inclusion of external resources for definitive diagnosis. Consequently, the gap between neurologists' normalized scores for differential (46.15%) and final diagnosis (70.93%) indicates that external resources likely enhance diagnostic accuracy beyond mere reliance on personal knowledge. In contrast, AI performed well in both differential (85%) and final diagnosis (88.24%), demonstrating capability in two distinct yet complementary diagnostic tasks: the substantiated formulation of a comprehensive array of hypotheses and the decisive synthesis of a conclusive diagnosis. Moreover, AI manifested perspicuity and cogency, with a significant record of four preferred answers selected by both evaluators. Nonetheless, neurology is a discipline reliant on collaborative, multidisciplinary problem-solving. In turn, comparing AI with neurologists' highest individual scores reveals a nuanced picture. In differential diagnosis, both the best neurologists and AI surpassed each other on two occasions. Remarkably, at least one neurologist achieved perfect scores for all final diagnoses, surpassing AI in two cases. Of course, using the highest individual scores as a proxy for collective intelligence does not fully capture the complex dynamics that influence the collective endorsement of individual contributions. In addition, AI's rapid generation of differential and final diagnoses, typically within one minute, contrasts the average times of 10 and 9 min, respectively, taken by neurologists. This disparity, particularly considering the neurologists' consistent reliance on external resources for final diagnoses, underscores the potential of AI as a time-efficient and resourceful asset in clinical practice, especially for individual clinical endeavors. It suggests workflow optimization potential, allowing physicians to focus more on patient interaction and treatment planning, though implementation frameworks must ensure AI remains a complementary tool that enhances rather than replaces clinical judgment.

Regarding AI's null scores, the first evaluator incurred a two-point deduction on a differential diagnosis challenge because the two least important diagnoses proposed were deemed aberrant (Supplementary S4). Otherwise, AI's answer would have received a score of 4 (out of 5 possible points based on this evaluator's scoring pattern for this question). Notably, the answer included the correct final diagnosis. This outcome likely resulted from the AI's prompt engineering strategy, inspired by 'surgical sieves' [29] to systematically evaluate various pathologies, leading to the listing of supernumerary diagnoses. Paradoxically, this reveals that while this algorithmic design enables the LLM to

emulate a human approach to diagnosis, it lacks innate human nuance in selecting relevant diagnoses—an intuition not emerging from our solution architecture and LLM’s foundational weights. Context-sensitive filtering and adaptive mechanisms of prompt engineering could better prioritize diagnoses based on clinical relevance while maintaining comprehensive differential coverage—addressing the gap between algorithmic thoroughness and human clinical intuition in diagnostic decision-making. The second evaluator attributed a null score on a final diagnosis challenge, applying a deliberately conservative scoring approach—i.e., recognizing only a unique option as granting points (Supplementary S4). While AI initially proposed the correct diagnosis in its differential, it did not retain it in its final diagnosis. Interestingly, the AI’s final diagnosis aligned with most neurologists (11 out of 13 human participants), leading to the highest number of null scores on a question record—questioning AI’s ability for original contributions.

4.1. Rationale

Since the public introduction of LLMs, many studies have compared AI and human performance. As this surge emphasizes the potential of generalist models, broadening accessibility for users of varying expertise, it also prompts scrutiny of their reliability in specialized contexts [30,31]. Indeed, these models were often applied to large datasets and within paradigms distant from real-world intricacies. On the other hand, seminal works have demonstrated promising results from large-scale projects conducted by major technology corporations, employing resource-intensive fine-tuning methods in sophisticated protocols [9,32]. However, many of these studies yielded outcomes that proved difficult to interpret [33]. We posited that a state-of-the-art foundational LLM, adeptly harnessed to a curated and indexed corpus of knowledge, can achieve domain-specific relevance in a complex clinical discipline. In parallel, we pursued controllability, verifiability, and scalability for mitigating influences and biases from the model’s pre-trained weights, enabling source tracking and advancing accessibility for clinicians and researchers. Testing this hypothesis, we conducted an in-depth, qualitative, and quantitative comparison of human and artificial diagnostic acumen, perspicuity, and cogency in a naturalistic setting. The findings support our rationale by confirming effective information synthesis from relevant selections of aptly cited sources and asserting the absence of hallucinatory content. In addition, this solution aims for versatility and accessibility. Relying on open-source technologies, it can be implemented on multiple foundational LLMs (e.g., GPT [34], Llama [35], and the Mistral series [36]). Its data-agnostic architecture accommodates numerous file formats for building a curated corpus.

4.2. Limitations

The limited cohort of neurologists and the select set of clinical cases do not fully represent the spectrum of neurological expertise or the complexities of clinical practice, potentially limiting the generalizability of the findings and constraining the solution’s broader applicability. Regarding the reliability of the AI-generated information, our two-fold approach—first categorizing the relevance of provided references and then applying binary classification of information accuracy—does not capture the subtle influences of the LLM’s constitutional weights. Indeed, the foundational LLM integral to our framework is inherently subject to biases [37,38], a byproduct of their data-driven training that lies beyond our control. This issue is exacerbated by proprietary and closed-source models [39] which also exposed to overfitting. Biases do not spare the corpus we employ [40], despite—and, in some aspects, because of—our control over its curation. Interestingly, despite access to diverse sources, the system’s exclusive retrieval from peer-reviewed textbooks suggests an inherent selection mechanism favoring authoritative academic content.


However, even rigorously vetted academic sources harbor inherent biases, emphasizing the critical importance of transparent source attribution that enables users to evaluate information provenance. This point and the curation and maintenance of the corpus itself were beyond the scope of this study. It raises intellectual property considerations and concerns about the quality and up-to-dateness of the information. As this research is nascent, we focused on solution development rather than its optimization. The system's performance can vary based on several factors, including the fine-tuned LLM for embeddings, the foundational LLM, their parameters, the prompt engineering strategies, and the structure and composition of the LTM and STM databases. The prototype corpus was assembled from raw, unprocessed textual data to uphold methodological neutrality and facilitate the agnostic, versatile approach we aimed to investigate. It remains plausible that text preprocessing techniques optimized for LLMs/NLP could enhance the system's performance. Also, we purposely deployed our solution on a limited but cohesive corpus of knowledge on clinical neurology. Further investigation is required to assess the solution's ability to manage a diverse and extensive corpus, including heterogeneous and conflicting sources, in a manner that effectively and transparently meets the end-user's needs and objectives. Recently, multimodal AI models [41,42], which can integrate natural language information with other sensory data such as images and audio, were introduced. This framework is not yet equipped to deploy these models. Finally, evaluating the solution's usability for non-expert users was beyond the scope of this study and will be the focus of subsequent research.

4.3. Perspectives

Humans have built, shared, and accessed knowledge in evolving ways. Transitioning from orality to literacy, and from analog to digital media, these evolutions have fundamentally shaped our comprehension of the world. Throughout these transitions, natural language remains central. Its adaptive and symbolic nature enables abstract thought and complex communication. With AI, the human ability for information integration and generation has been challenged—LLMs' prowess for NLP is humbling in this regard. It elucidates the singular role of natural language in structuring knowledge derived from various modalities, media, and agents. This prompts reconsideration of complex tasks once thought uniquely human. While LLMs excel at information-intensive tasks, they lack general reasoning capabilities [43–47] and are grounded in human-derived data. This raises questions about their efficacy in scenarios requiring original thinking or high-order cognition and about potential bias propagation [48].

LLMs indicate a new phase of human–machine integrative intelligence, with profound implications for cognition and knowledge. In medicine, LLMs can complement clinicians' experiential understanding by integrating patient histories, physical examinations, and test findings with vast medical knowledge. This emphasizes the need for human oversight and contextual interpretation.

Explainable AI (XAI) and specialized, purpose-built interfaces (Figure 6) will be crucial for ethical integration and utility evaluation in high-stakes fields. It is imperative to involve and empower clinicians and scientists in LLM-driven applications and research, developing accessible tools and consensus-based standards aligned with medical needs and priorities. This approach ensures that as our tools become more sophisticated, they remain controllable and serve our purposes rather than obfuscate them.



Username

Password

Select your Topic
vns_review

Select Source

kahlow2013_arti...

seki2014_article

mcgregor2005_a...

spuck2008_article

cantarín-extrem...

ardesch2007_art...

Additional Settings
Model
gpt-4

Temperature
0.15

Contexts
10

Token Length
1000

Base Prompt
audience or academic neurologists. At the end of your response, provide a markdown-formatted bibliography named 'Bibliography' that lists all sources cited, including specific pages, DOI, and the APA citation.

Save Settings


 What are the indications of right-sided vagus nerve stimulation?

 Right-sided vagus nerve stimulation (R-VNS) has been used as a treatment option for refractory epilepsy. This approach has been particularly beneficial in cases where patients have shown significant improvement with left-sided vagus nerve stimulation (L-VNS), but the devices had to be removed due to various reasons (McGregor et al., 2005, p. 1). However, it's important to note that traditionally, the left vagus nerve has been preferred for stimulation due to concerns about increased risk of cardiac symptoms with right-sided stimulation. The right vagus nerve provides innervation to the atrioventricular (AV) node, and its stimulation has been associated with bradycardia (Cantarín-Extremera et al., 2016, p. 3).

Bibliography:

- McGregor, A., Wheless, J., Baumgartner, J., & Bettis, D. (2005). Right-sided Vagus Nerve Stimulation as a Treatment for Refractory Epilepsy in Humans. In *Epilepsia* (Vol. 46, Issue 1, pp. 91–96). Wiley. <https://doi.org/10.1111/j.0013-9580.2005.16404.x>
- Cantarín-Extremera, V., Ruíz-Falcó-Rojas, M.-L., Tamariz-Martel-Moreno, A., García-Fernández, M., Duat-Rodríguez, A., & Rivero-Martín, B. (2016). Late-onset periodic bradycardia during vagus nerve stimulation in a pediatric patient. A new case and review of the literature. In *European Journal of Paediatric Neurology* (Vol. 20, Issue 4, pp. 678–683). Elsevier BV. <https://doi.org/10.1016/j.ejpn.2016.02.014>

What do you want to know?

Figure 6. Web purpose-built interface in basic mode provides the ability to select a corpus, define a knowledge base with specific sources, choose the foundational LLM deployed, parameterize its behavior, and set base prompt engineering.

5. Conclusions

In our comparative analysis, a specialized LLM demonstrated superior diagnostic performance against practicing neurologists in complex neurological cases, highlighting its potential as a time-efficient clinical asset. The solution's architecture employing RAG on a curated corpus achieved domain-specific relevance while maintaining verifiability through source tracking. Our findings suggest a promising trajectory for human–machine integrative intelligence in healthcare. This research underscores the importance of developing accessible, transparent AI tools that complement rather than replace clinical expertise. As we advance these technologies, maintaining human oversight and aligning development with medical priorities will ensure AI systems remain valuable tools that enhance rather than diminish the essential human dimensions of healthcare.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/brainsci15040347/s1>, Supplementary S1. Clinical cases; Supplementary S2. Scoring sheets; Supplementary S3. AI Sourced Replies; Supplementary S4. AI's null scores; Figure S1: Normal Q-Q plot; Figure S2: Plot of residuals. References [49–53] are cited in the supplementary materials.

Author Contributions: All authors have made substantial contributions to the conception and design of the study, acquisition of data, or analysis and interpretation of data. Specifically, the contributions are as follows: S.B. (Sami Barrit) and N.T. contributed to the conception and design of the study, data acquisition, analysis, interpretation, drafting the manuscript, and the final approval of the version to be published. A.M. (Aurelien Mazeraud), S.L. and R.C. participated in the design of the study, data analysis and interpretation, the critical revision of the manuscript for important intellectual content, and the final approval of the version to be published. S.B. (Sebastien Boulogne), J.B., T.C. (Timothée Carette), T.C. (Thibault Carron), M.A.B., B.D., E.D., H.K., A.M. (Adil Maarouf), S.M.S., S.R., A.R., S.H., V.H., V.T. and T.M. contributed to data acquisition, data analysis, interpretation, the review of the manuscript, and the final approval of the version to be published. A.N. and S.E.H. contributed to data analysis and interpretation, manuscript revision for intellectual content, and the final approval of the version to be published. N.M. and J.R.M. contributed to data analysis and provided critical revisions and final approval of the version to be published. Additionally, all authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Computing resources were provided by Science (New York, NY, US) through a nonprofit program supporting open and decentralized science.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Anonymized data not published within this article will be made available by request to the corresponding author from any qualified investigator upon reasonable request.

Conflicts of Interest: All authors affiliated with Science (New York, NY, US) declare interest in creating an organization aimed at developing open and decentralized science and facilitating AI's usage in medicine. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
LLM	Large language model
RAG	Retrieval-augmented generation

References

1. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef] [PubMed]
2. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Zhang, J.; et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* **2021**, *2*, 100179. [CrossRef] [PubMed]
3. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *Preprint* **2018**. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 March 2023).
4. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
5. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; McGrew, B.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
6. Beam, A.L.; Drazen, J.M.; Kohane, I.S.; Leong, T.Y.; Manrai, A.K.; Rubin, E.J. Artificial Intelligence in Medicine. *N. Engl. J. Med.* **2023**, *388*, 1220–1221. [CrossRef]
7. Ling, C.; Zhao, X.; Lu, J.; Deng, C.; Zheng, C.; Wang, J.; Chowdhury, T.; Li, Y.; Cui, H.; Zhao, L.; et al. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *arXiv* **2023**, arXiv:2305.18703. [CrossRef]
8. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:1906.02243. [CrossRef]
9. Singhal, K.; Azizi, S.; Tu, T.; Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Natarajan, V.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef]
10. Lipton, Z.C. The Mythos of Model Interpretability. *Queue* **2018**, *16*, 31–57. [CrossRef]
11. Huang, K.; Altaosaar, J.; Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv* **2019**, arXiv:1904.05342. [CrossRef]
12. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts (Version 3). *arXiv* **2023**, arXiv:2307.03172. [CrossRef]
13. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**, arXiv:2005.11401. [CrossRef]
14. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
15. Pokorný, J. NoSQL databases. In Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, Ho Chi Minh City, Vietnam, 5–7 December 2011. [CrossRef]
16. Taipalus, T. Vector database management systems: Fundamental concepts, use-cases, and current challenges. *arXiv* **2023**, arXiv:2309.11322. [CrossRef]
17. Han, M.H. *Adams and Victor's Principles of Neurology*; American Association of Neuropathologists, Inc.: Littleton, CO, USA, 2009.
18. Brazis, P.W.; Masdeu, J.C.; Biller, J. *Localization in Clinical Neurology*, 6th ed.; Wolters Kluwer Health Adis (ESP): Waltham, MA, USA, 2012; pp. 1–668.
19. Jankovic, J.; Mazziotta, J.C.; Pomeroy, S.L.; Newman, N.J. *Bradley's Neurology in Clinical Practice*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2021.
20. Cooper, P.E. Cooper P.E. DeJong's The Neurologic Examination. 2005. Sixth edition. By William W. Campbell. Published by Lippincott, Williams & Wilkins. 671 pages. C\$140 approx. *Can. J. Neurol. Sci.* **2017**, *32*, 558. [CrossRef]
21. Rowland, L.P.; Pedley, T.A.; Merritt, H.H. *Merritt's Neurology*; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2010.
22. Edition MMP. Neurologic Disorders. 2023. Available online: <https://www.msmanuals.com/professional/neurologic-disorders> (accessed on 25 September 2023).
23. Wikipedia. Category: Neurological Disorders. 2023. Available online: https://en.wikipedia.org/wiki/Category:Neurological_disorders_%E2%80%8C (accessed on 25 September 2023).
24. Lun, R.; Niznick, N.; Padmore, R.; Mack, J.; Shamy, M.; Stotts, G.; Blacquiére, D. Clinical Reasoning: Recurrent strokes secondary to unknown vasculopathy. *Neurology* **2020**, *94*, e2396–e2401. [CrossRef]
25. Francis, A.W.; Kiernan, C.L.; Huvard, M.J.; Vargas, A.; Zeidman, L.A.; Moss, H.E. Clinical Reasoning: An unusual diagnostic triad. Susac syndrome, or retinocochleocerebral vasculopathy. *Neurology* **2015**, *85*, e17–e21. [CrossRef]
26. Choi, J.H.; Wallach, A.I.; Rosales, D.; Margiewicz, S.E.; Belmont, H.M.; Lucchinetti, C.F.; Minen, M.T. Clinical Reasoning: A 50-year-old woman with SLE and a tumefactive lesion. *Neurology* **2017**, *89*, e140–e145. [CrossRef]
27. Harada, Y.; Elkhider, H.; Masangkay, N.; Lotia, M. Clinical Reasoning: A 65-year-old man with asymmetric weakness and paresthesias. *Neurology* **2019**, *93*, 856–861. [CrossRef]
28. McIntosh, P.; Scott, B. Clinical Reasoning: A 55-Year-Old Man with Odd Behavior and Abnormal Movements. *Neurology* **2021**, *97*, 1090–1093. [CrossRef]
29. Chai, J.; Evans, L.; Hughes, T. Diagnostic aids: The Surgical Sieve revisited. *Clin Teach.* **2017**, *14*, 263–267. [CrossRef]

30. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Tseng, V.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef]
31. Schubert, M.C.; Wick, W.; Venkataramani, V. Performance of Large Language Models on a Neurology Board-Style Examination. *JAMA Netw. Open* **2023**, *6*, e2346721. [CrossRef] [PubMed]
32. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv* **2023**, arXiv:2305.09617. [CrossRef]
33. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
34. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
35. Touvron, H.; Martin, L.; Stone, K.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
36. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
37. Li, Y.; Du, M.; Song, R.; Wang, X.; Wang, Y. A Survey on Fairness in Large Language Models. *arXiv* **2023**, arXiv:2308.10149. [CrossRef]
38. Wu, M.; Fikri Aji, A. Style Over Substance: Evaluation Biases for Large Language Models. *arXiv* **2023**, arXiv:2307.03025. [CrossRef]
39. Sanderson, K. GPT-4 is here: What scientists think. *arXiv* **2023**, 615, 773. [CrossRef]
40. Louie, P.; Wilkes, R. Representations of race and skin tone in medical textbook imagery. *Soc. Sci. Med.* **2018**, *202*, 38–42. [CrossRef] [PubMed]
41. Belyaeva, A.; Cosentino, J.; Hormozdiari, F.; Eswaran, K.; Shetty, S.; Corrado, G.; Carroll, A.; McLean, C.Y.; Furlotte, N.A. Multimodal LLMs for health grounded in individual-specific data. *arXiv* **2023**, arXiv:2307.09018. [CrossRef]
42. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv* **2023**, arXiv:2306.09093. [CrossRef]
43. Chollet, F. On the Measure of Intelligence. *arXiv* **2019**, arXiv:1911.01547. [CrossRef]
44. Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Cooper Stickland, A.; Korbak, T.; Evans, O. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. *arXiv* **2023**, arXiv:2309.12288. [CrossRef]
45. Dziri, N.; Lu, X.; Sclar, M.; Li, X.L.; Jiang, L.; Lin, B.Y.; Welleck, S.; West, P.; Bhagavatula, C.; Le Bras, R.; et al. Faith and Fate: Limits of Transformers on Compositionality. *arXiv* **2023**, arXiv:2305.18654. [CrossRef]
46. McCoy, R.T.; Yao, S.; Friedman, D.; Hardy, M.; Griffiths, T.L. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv* **2023**, arXiv:2309.13638. [CrossRef]
47. Mitchell, M.; Palmarini, A.B.; Moskvichev, A. Comparing Humans, GPT-4, and GPT-4V On Abstraction and Reasoning Tasks. *arXiv* **2023**, arXiv:2311.09247. [CrossRef]
48. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey. *arXiv* **2023**, arXiv:2309.00770. Available online: <https://ui.adsabs.harvard.edu/abs/2023arXiv230900770G> (accessed on 1 September 2023). [CrossRef]
49. Daroff, R.B.; Jankovic, J.; Mazziotta, J.C.; Pomeroy, S.L.; Bradley, W.G. *Bradley’s Neurology in Clinical Practice*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 149, 237, 304, 334, 338, 564, 569, 570, 1051, 1061, 1067, 1075, 1181, 1192, 1223, 1256, 1257, 1294, 1361, 1828, 1890, 2243, 2312, 2323–2325, 2330, 2337, 2339, 2341, ISBN 0323339166.
50. Rowland, L.P.; Pedley, T.A.; Merritt, H.H. *Merritt’s Neurology*; Wolters Kluwer: Alphen aan den Rijn, The Netherlands, 2016; pp. 854, 690, 1180, 1348, 1445, 1472, ISBN 145119336X.
51. Ferreri, A.J.; Campo, E.; Seymour, J.F.; Willemze, R.; Ilariucci, F.; Ambrosetti, A.; Zucca, E.; Rossi, G.; López-Guillermo, A.; Pavlovsky, M.A.; et al. Intravascular lymphoma: Clinical presentation, natural history, management and prognostic factors in a series of 38 cases, with special emphasis on the ‘cutaneous variant’. *Br. J. Haematol.* **2004**, *127*, 173–183. [CrossRef] [PubMed]
52. Ropper, A.; Samuels, M.; Klein, J. *Adams and Victor’s Principles of Neurology*, 10th ed.; McGraw-Hill: New York, NY, USA, 2014; pp. 889, 1224, 1543, 2032, ISBN 978-0071794794.
53. Jung, H.H.; Danek, A.; Walker, R.H. Neuroacanthocytosis Syndromes. *Orphanet J. Rare Dis.* **2011**, *6*, 68. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.