

ORIGINAL RESEARCH

Open access

Evaluating Large Language Models for Symptom-Based Medical Diagnosis: A Comparative Performance Analysis

Rafael Mendoza^{1*}, Isabela Cruz¹, Carlos Lopez²

Abstract

This research undertook an assessment of Large Language Models (LLMs) within the domain of clinical diagnostics, specifically investigating their capacity to interpret symptom-centered inputs and generate precise clinical assessments. The investigation focused on evaluating several prominent architectures—including GPT-4, GPT-4o, Gemini, o1 Preview, and GPT-3.5—to gauge their proficiency in identifying clinical conditions solely from presented symptomatology. To guarantee clinical validity and appropriateness, symptom profiles were compiled from verified medical literature. Each architecture underwent testing within standardized environments to quantify diagnostic accuracy, precision, recall, and underlying logical processing. Specialized clinical scenarios were structured to examine performance boundaries across both routine and critical diagnostic conditions. The findings indicated that GPT-4 delivered the highest overall diagnostic accuracy, displaying a strong synergy with established medical logic. Conversely, Gemini demonstrated superior performance in high-stakes environments where exact decision-making is vital. Both GPT-4o and o1 Preview presented equitable metrics, maintaining a stable balance between precision and recall while managing real-time diagnostic scenarios. While possessing less advanced capabilities, GPT-3.5 remained functional for standard diagnostic duties. Ultimately, this investigation outlines the current capabilities and constraints of utilizing LLMs for clinical determination. Even though architectures like GPT-4 and Gemini show significant capability, critical barriers regarding data confidentiality, ethical protocols, and training bias correction must still be overcome. These insights offer strategic directions for the ethical incorporation of LLMs into diagnostic frameworks to optimize patient care outcomes.

Keywords Natural language processing (NLP), Healthcare, Digital health, AI, Large language models, Medical diagnostics

*Correspondence:

Rafael Mendoza
rafael.mendoza@unam.mx

¹ Department of Clinical AI and Large Language Models, Faculty of Medicine, National Autonomous University of Mexico, Mexico City, Mexico

² Department of Symptom-Based Diagnosis and LLM Evaluation, Faculty of Engineering, Monterrey Institute of Technology, Monterrey, Mexico

Introduction

Progressions in the field of Large Language Models (LLMs) have fundamentally altered natural language processing, introducing novel avenues for clinical operations and medical care delivery [1-7]. Within medical environments, LLMs are increasingly viewed as viable instruments for boosting diagnostic precision and optimizing clinical communication through the interpretation and production of human-like language [8, 9]. In light of heightened medical demands and the necessity for streamlined patient triage, these computational systems could automate structural components of diagnostic pipelines, expanding the availability of clinical resources while mitigating clinician burnout [10]. LLMs have previously validated their

utility in information-heavy domains such as radiology and pathology, where the rapid analysis of diagnostic media assists in early-stage pathology identification and therapeutic structuring [11].

In addition to diagnostic tasks, LLMs assist in patient engagement initiatives by providing customized medical orientations and symptom triage, which can strengthen patient trust and rapport. Through the evaluation of heterogeneous information streams—such as longitudinal case records, diagnostic imaging, and telemetry from biometric wearables—these systems empower healthcare practitioners with data-driven clinical insights [12]. These implementations also reach into psychiatric care, where speech and behavioral patterns can be assessed, chronic care coordination where ongoing patient telemetry is tracked, and ophthalmology through the early identification of disorders such as diabetic retinopathy [13].

Nonetheless, deploying LLMs within clinical settings introduces substantial hurdles, particularly regarding patient data confidentiality, model transparency, and the ethical dimensions of automated clinical choices. Compliance with statutory data protection mandates, such as HIPAA regulations, remains a non-negotiable prerequisite [14]. Furthermore, intrinsic algorithmic biases within LLMs must be systematically corrected to prevent detrimental impacts on patient management and diagnostic validity [15]. Resolving these technical, ethical, and legislative challenges is imperative for the secure adoption of LLMs in active clinical practice.

This research provides a structured evaluation of the diagnostic proficiencies of five cutting-edge Large Language Models (LLMs)—specifically GPT-4, GPT-4o, Gemini, o1 Preview, and GPT-3.5—utilizing a standardized testing matrix to guarantee equity and uniformity during cross-model benchmarking. The empirical data highlights the outstanding diagnostic precision of GPT-4, facilitated by its comprehensive training across varied informational domains, which aids its management of complex clinical presentations. Gemini proves uniquely effective in high-acuity contexts by emphasizing precision metrics and reducing false-positive rates, whereas GPT-4o and o1 Preview demonstrate a balanced matrix, pairing real-time operational readiness with trustworthy precision. Although operating with reduced computational depth, GPT-3.5 remains practical for baseline diagnostic triage and resource-constrained environments. This investigation also draws attention to major limitations hindering these systems, such as struggles when translating vague or nuanced symptom expressions, the imperative for airtight privacy firewalls, and the risk of reinforcing biases embedded within training corpora. These insights expand current perspectives on the diagnostic utility of LLMs and lay a groundwork for their ethical introduction into day-to-day clinical workflows.

Related Work

Advancements in LLM capabilities for healthcare applications

The adoption of Large Language Models (LLMs) within medical systems is reshaping diagnostic frameworks and patient management [16]. These architectures refine both the speed and exactness of diagnostic operations, particularly in medical specialties like radiology and pathology that rely on granular examinations and early-stage disease detection [11]. Furthermore, LLMs optimize the patient experience by facilitating individualized health dialogues and symptom triage, which serves to build patient confidence and engagement. Beyond primary diagnostics, these computational tools process multi-source data repositories—encompassing historical medical records, diagnostic imaging, and streaming wearable telemetry—to guide healthcare professionals toward well-informed clinical actions [12]. Along these lines, Meskó *et al.* [17] underscore the capacity of architectures like GPT-4 to manage massive datasets and synthesize human-like text, establishing an integrated medical methodology that fuses textual, visual, and sensor-based telemetry channels. This multi-modal consolidation allows for holistic patient evaluations and ultimately elevates the quality of medical care. Paralleling these findings, alternative investigations [13, 18] have shown that LLMs can expand into behavioral health domains by parsing behavioral markers to forecast clinical pathways, offering a supportive resource for psychiatric specialists designing customized intervention strategies.

Challenges in Integrating LLMs into healthcare

The incorporation of generative AI and LLMs into medical infrastructures brings about multifaceted and intertwined challenges [19–26]. Yu et al. [27] emphasize the imperative for strict data security protocols, targeted parameter fine-tuning, and comprehensive deployment blueprints to guarantee that AI utilization mirrors clinical demands without sacrificing system safety or operational throughput. They champion the paradigm of participatory co-design—uniting medical practitioners and patients in the AI engineering lifecycle—to ensure software utilities match clinical prerequisites while preserving data integrity and patient confidentiality [14]. Singh et al. [28] focus on psychological vulnerabilities, notably the cognitive biases embedded in LLMs that can trigger overconfident or under-calculated diagnostic projections. Such distortions risk causing flawed clinical choices, highlights the critical necessity for diagnostic features that evaluate and calibrate AI confidence thresholds. Additionally, Ullah et al. [29] detail architectural limitations within diagnostic disciplines, highlighting constraints surrounding contextual awareness and explainability in areas like digital pathology. The opaque “black-box” architecture characteristic of LLMs, paired with structural biases within foundational training sets, complicates their clinical validation and institutional reliability [30]. Consequently, the practical deployment of these technologies encounters serious bottlenecks, including the clinical consistency of AI tools in active care environments, the requirement for extensive clinical trials, and persistent anxieties regarding the privacy and protection of sensitive patient profiles [31].

LLMs in clinical trials and patient monitoring

LLMs have demonstrated considerable capacity for optimizing clinical trials and continuous patient tracking by upgrading data-processing methodologies and predictive metrics. Yuan et al. [32] illustrate the ways in which LLMs refine clinical trial workflows via automated patient-trial matching that leverages Electronic Health Records (EHRs). Their privacy-preserving data-augmentation framework yielded a 7.32% improvement in matching accuracy, accelerating both the velocity and exactness of cohort selection for investigational trials, which subsequently accelerates therapeutic discoveries. Jin et al. [33] and Kim et al. [34] investigate the Health-LLM framework, which implements sophisticated algorithms to evaluate streaming data from wearable telemetry in real time. This architecture yields customized health projections that dynamically adapt to shifts in patient status, particularly for managing chronic pathologies. These inquiries highlight the potential of LLMs to facilitate ongoing health tracking and contribute to preventative medicine by promoting adherence to recommended therapeutic routines. Xu et al. [35] examine specialized Mental-LLM utilities that parse linguistics in real time to forecast psychiatric trends, providing valuable diagnostic insights for longitudinal monitoring in mental health cases. Resolving these operational difficulties requires continuous progress in AI architectures to optimize their precision, dependability, and generalizability across varied clinical settings [36, 37].

Applications and limitations of LLMs in diagnostics

The deployment of LLMs for clinical determination highlights both their analytical potential and systemic bottlenecks. Meskó et al. [17] and Kusa et al. [12] offer foundational insights regarding the capacity of LLMs to synthesize multimodal data streams, an approach that can refine diagnostic precision and enhance clinical outcomes. Conversely, Kusa et al. also emphasize vulnerabilities associated with LLM sensitivity to subtle variations in user prompts, which can cause inconsistencies in diagnostic conclusions. Such variations frequently originate from disparate symptom descriptions and deeply held patient assumptions, highlighting the necessity for adaptive frameworks that accommodate these individual variations to prevent diagnostic errors. Furthermore, an investigation by Abbasian et al. [23] analyzed how LLM-driven conversational medical agents upgraded the patient experience during diagnostic triage. Nonetheless, the dependence of LLMs on high-quality intake information continues to be a major obstacle, as erratic input data can undermine model efficiency and trustworthiness [27]. The perspectives gained from these investigations are vital for mapping the diagnostic role of LLMs and establishing the protocols required to manage prompt sensitivity and input variability.

Ethical and technical considerations for LLM deployment

The implementation of LLMs within healthcare demands rigorous adherence to technical and ethical protocols. Montagna *et al.* [11] outline the difficulties of deploying LLM-based conversational agents for chronic disease oversight, particularly inside decentralized frameworks where information privacy and data security are paramount. Their proposed operational framework for chronic care coordination accommodates varied clinical pathologies while complying with stringent data-protection mandates; however, its widespread adoption is hindered by reliability constraints and the requirement for exhaustive clinical validation trials. Correspondingly, Kim *et al.* [34] delineate the technical complexities involved in processing multimodal data arrays in real time, particularly within niche medical subspecialties. Overcoming these barriers requires simultaneous advancements in AI engineering and regulatory governance. Humphrey *et al.* [14] and Yuan *et al.* [32] advocate for ethical AI methodologies and strict statutory alignment to protect patient confidentiality and data integrity. Satisfying these criteria will be crucial for optimizing the clinical utility of LLMs while ensuring a responsible, safe, and secure integration process.

Materials and Methods

Research strategy

This investigation employed a structured methodology to gauge the diagnostic proficiency of Large Language Models (LLMs) through symptom-oriented textual prompts extracted from an engineered dataset (Figure 1). Each LLM received an identical sequence of prompts to ensure that any discrepancies in the generated outputs could be attributed exclusively to the distinct interpretive frameworks of the respective models. This standardized protocol ensured uniform comparative analysis across all systems, maintaining absolute consistency in prompt delivery and output logging while eliminating confounding variables associated with divergent input structures or varying assessment protocols. The compiled dataset was sourced from authoritative medical institutions, comprising public clinical guidelines and validated disease-symptom matrices. Every prompt was formulated to mirror authentic clinical encounters by blending commonly co-occurring symptoms, such as “cough, mild headache, sneezing”. Structuring the dataset around diverse symptom portfolios guaranteed that the models were challenged across a broad spectrum of diagnostic difficulty.

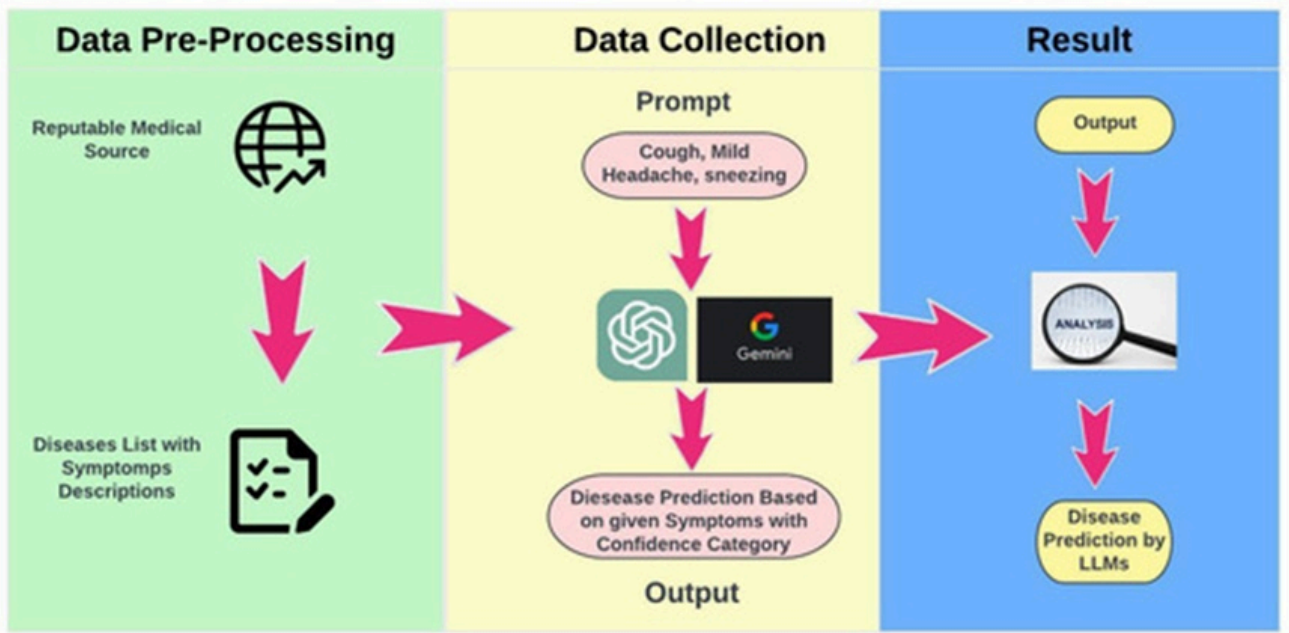


Figure 1. Research Strategy.

Following the collection of model-generated predictions, every output was systematically cross-referenced with the verified gold-standard diagnosis linked to that specific symptom profile. This assessment utilized a zero-shot diagnostic

paradigm, meaning the architectures generated clinical predictions without any prior fine-tuning or domain-specific adaptations. Performance was quantified using standard parameters, including precision, recall, and F1 scores, which provided an all-inclusive assessment of diagnostic exactness. This systematic paradigm not only uncovered individual architectural advantages and structural deficits but also provided a durable framework for assessing the general dependability and clinical relevance of LLMs in medicine. The uniform execution of this methodology minimized algorithmic bias and external confounding factors, ensuring an objective and transparent analysis of each model's capabilities. These insights help advance the understanding of LLM utility in medical assessment and offer a baseline for subsequent research targeting highly complex clinical settings and multimodal data integration.

Description of the LLMs evaluated

This section outlines the specific Large Language Models (LLMs) analyzed in this research—namely Gemini, GPT-3.5, GPT-4, o1 Preview, and GPT-4o. Each architecture presents distinct advantages in synthesizing clinically relevant text and executing processes like clinical forecasting, diagnostic triage, and producing data-driven strategies to support patient health restoration and maintenance. These LLMs are highly accessible, regularly utilized by laypersons, and represent crucial technologies requiring rapid benchmark evaluations to establish their viability and efficacy for clinical deployment and health sciences research. Every individual model features specialized computational dynamics for interacting with clinical information arrays, helping to improve general diagnostic reliability across healthcare environments [12].

Gemini: Gemini represents a major technological leap in LLM development, particularly through its tailored engineering for domain-specific tasks, including medicine. The foundational architecture of Gemini is optimized to facilitate granular comprehension and text generation within specialized sectors, rendering it a highly useful instrument for duties such as clinical diagnostics and medical inquiries where exactness and precision are paramount. Gemini's capacity to process and reason across multimodal informational inputs highlights its potential to revolutionize medical data workflows, setting a new benchmark for healthcare AI [7].

GPT-3.5: Serving as a baseline comparison model, GPT-3.5—the immediate predecessor to GPT-4—presents capable functionalities in language translation and text generation, albeit with slightly reduced proficiency relative to subsequent generations. GPT-3.5 is utilized here as a comparative baseline to measure structural advancements in LLM technology and its practical utility in medical assessment. Despite its legacy status, GPT-3.5 attained a notable 53 percent accuracy rate when tested on the Medical Knowledge Self-Assessment Program, demonstrating its core capability to interpret and address clinical and medical queries, thereby validating its utility for basic diagnostic tasks [5].

GPT-4: Engineered by OpenAI, GPT-4 represents a leading standard in linguistic comprehension and text generation, configured to unpack complex informational requests. Its underlying architecture is specifically calibrated to judge diagnostic exactness based on textual symptom profiles. GPT-4 achieved a 75 percent accuracy rate on the Medical Knowledge Self-Assessment Program, highlighting its sophisticated comprehension of complex clinical inquiries and emphasizing its utility in optimizing diagnostic accuracy derived from symptom narratives [5].

o1 Preview: As a specialized release within the GPT lineage, o1 Preview prioritizes elevated diagnostic exactness alongside real-time processing speeds for medical use cases. Its underlying architecture is tuned to manage intricate analytical requests rapidly, making it exceptionally useful in clinical environments that demand time-sensitive diagnostic assistance. Building upon the core strengths of the GPT-4 framework, o1 Preview is engineered for smooth deployment within fast-paced clinical workflows, assisting medical staff in reaching rapid and precise diagnostic determinations, which marks it as a highly practical utility in contemporary medical operations [5].

GPT-4o: Configured as an optimized iteration of the GPT-4 baseline, GPT-4o prioritizes balanced, real-time diagnostic performance, especially inside active clinical workflows. This system refines the core competencies of GPT-4, adapting them to clinical environments that require immediate, accurate parsing of patient symptom data. OpenAI engineered GPT-4o with an emphasis on diagnostic precision while maintaining fluid integration into medical networks that depend on rapid decision-making. By emphasizing real-time execution, GPT-4o supports medical practitioners in making well-informed clinical judgments efficiently and reliably [5].

Data collection methods

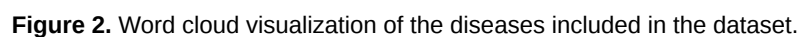
The research dataset was engineered around 50 distinct pathological conditions, with each condition evaluated across the selected Large Language Models (LLMs): Gemini, GPT-3.5, GPT-4, GPT-4o, and the o1 preview model. This structured paradigm generated a cumulative total of 250 standalone analytical trials (50 diseases × 5 models), as detailed in Table 1.

Table 1. Distribution of dataset evaluations across models.

Model	Diseases Evaluated per Model
Gemini	50
GPT-3.5	50
GPT-4	50
o1 Preview Model	50
GPT-4o	50
Total	250 total evaluations

Source and selection of diseases: Symptom profiles for the chosen medical conditions were harvested from authoritative medical repositories, including the Centers for Disease Control and Prevention (CDC) [38], the World Health Organization (WHO) [39], the Mayo Clinic [40], the Cleveland Clinic [41], and the Johns Hopkins Hospital [42]. For every clinical condition, an exhaustive registry of symptoms was compiled, establishing the baseline framework for the generation of the diagnostic prompts.

Dataset and disease selection: The testing matrix incorporated 50 commonplace clinical pathologies (Figure 2), encompassing conditions such as seasonal allergies, the common cold, and foodborne illnesses, which are frequently encountered within everyday primary care practices. Every disease profile was processed by all five models (Gemini, GPT-3.5, GPT-4, o1 Preview, and GPT-4o), ensuring strict uniformity across the model comparison matrix and generating 250 total data points for analytical review. This dataset targeted widely documented clinical markers to judge the models' capacity to yield precise diagnostic insights. The underlying dataset is publicly available via GitHub at: https://github.com/gkgupta11k/Health_LLM_Research_Dataset (accessed on 13 January 2025).



Diagnostic procedure: For every chosen pathological condition, a comprehensive profiling of clinical manifestations was compiled, establishing the basis for the diagnostic prompts. These prompts instructed the LLMs to forecast the specific condition derived from the symptom profiles and to generate an accompanying confidence metric for each output. To maintain methodological consistency and cross-model comparability, identical prompts were deployed uniformly across all evaluated architectures. Following the collection of the model-generated predictions, the outputs underwent manual verification to determine the accuracy and dependability of the clinical determinations, establishing the empirical foundation for the study's conclusions.

Within this study, the collective prompting phase involved 250 standalone iterations (reflecting the combination of 50 distinct diseases evaluated across 5 separate LLMs). Every text prompt supplied the model with a specific cluster of clinical manifestations and requested a single-word diagnostic determination along with an estimated confidence level, explicitly omitting any requirement for step-by-step rationale. This architectural design adhered to a zero-shot learning paradigm, wherein the models were required to formulate diagnostic predictions based exclusively on the provided symptoms without the aid of contextual examples or sequential chain-of-thought processing. This framework directly tested each model's intrinsic capacity to interpret clinical data and independently isolate the most probable pathology, simulating a direct diagnostic task common in clinical settings where minimal extraneous context is available.

Prompt for models:

The following dialogue presents a prompt used to test the diagnostic capabilities of various language models.

Based on these symptoms: [Symptoms], identify the only one disease based on the symptoms that match closely. Provide me confidence level—Low, Medium, High—to each, based on how closely the symptoms align with the diseases you have predicted based on the Symptoms. No explanation needed, provide me exact one-word disease name.

Evaluation metrics for diagnosing diseases through LLMs

The diagnostic efficacy of Language Learning Models (LLMs) in identifying clinical conditions from textual symptom descriptions was quantified through a systematic, multi-stage evaluation protocol. This analytical framework leveraged precision, recall, and the F1 score—statistical metrics widely recognized for their capacity to deliver a comprehensive perspective on predictive model accuracy by measuring correct identifications against the omission of valid clinical conditions.

In this investigation, we analyzed the outputs generated by the LLMs for every dataset entry, systematically categorizing each response according to its diagnostic validity. The classification criteria were defined as follows:

- **True positive (TP):** Occurrences where the LLM correctly identified the underlying pathology, demonstrating the model's capacity to accurately map clinical symptom profiles to the corresponding medical diagnosis.
- **False positive (FP):** Occurrences where the LLM generated an incorrect diagnostic determination, attributing an erroneous condition to a symptom profile that did not align with the ground-truth pathology, thereby overestimating the model's diagnostic accuracy.
- **False negative (FN):** Occurrences where the LLM either produced an alternative diagnosis instead of the actual underlying condition or failed to generate any diagnostic output whatsoever, thereby underestimating the model's diagnostic sensitivity.

We subsequently calculated the primary evaluation metrics based on the scoring logic assigned to each individual prediction. If the model's diagnostic output matched the ground-truth condition, it was awarded 1 point within the true positive (TP) category. If the model generated a diagnosis that deviated from the actual condition (an erroneous prediction), it was assigned 1 point within the false positive (FP) category. Lastly, if the system failed to isolate the correct pathology—either by generating an incorrect condition or by failing to return any response—it was assigned 1 point within the false negative (FN) category. Following the conclusion of the evaluation phase for each respective model, the cumulative TP, FP, and FN scores were aggregated. These sums were then utilized within the following mathematical expressions to compute precision, recall, and the F1 score:

- **Precision:** This metric quantifies the exactness of the model's positive diagnostic assertions (representing the proportion of TP observations relative to the sum of all positive diagnoses generated by the system), offering an indicator of the model's accuracy in disease differentiation.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

- **Recall:** This metric measures the model's capacity to capture all relevant clinical instances (representing the ratio of TP observations relative to all actual positive conditions present within the dataset), providing a diagnostic index of the model's comprehensiveness in disease detection.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

- **F1 Score:** This metric functions as a balanced, harmonic synthesis of both precision and recall, serving as an optimal diagnostic indicator when both statistical parameters carry equal analytical weight. It is computed as the harmonic mean of precision and recall, yielding a singular, unified metric of the model's absolute diagnostic performance.

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

The deployment of these statistical metrics facilitated a rigorous, multi-faceted evaluation of the LLMs' diagnostic proficiencies, yielding granular insights into the exactness of correct clinical determinations alongside the general efficacy of the models in pathological identification.

Results and Discussion

Overview of findings

This thorough investigation scrutinized the diagnostic proficiencies of five cutting-edge Large Language Models (LLMs)—Gemini, GPT-3.5, GPT-4, o1 Preview, and GPT-4o—focusing on their capacity to parse and interpret intricate symptom profiles to generate valid medical assessments. The resulting data, mapped out in **Figure 3**, exposed pronounced disparities in diagnostic competency and accuracy across the architectures, offering vital insights into their clinical viability. GPT-4 distinguished itself as the premier system in this evaluation, exhibiting exceptional diagnostic precision. The superior performance of this architecture stems from its extensive pre-training on a vast corpus of medical literature and clinical records, which endows it with a nuanced comprehension of pathological indicators. GPT-4's capacity to reliably isolate specific diseases from textual symptom lists reflects its highly advanced algorithmic framework and refined data processing capabilities. By establishing a new benchmark for AI-driven clinical determination, it proves to be a durable asset capable of significantly modernizing diagnostic and therapeutic workflows.

Closely trailing the top tier, o1 Preview also displayed remarkable diagnostic capabilities, yielding a precision of 0.93 alongside a recall of 0.91. This output rendered it highly competitive with GPT-4, showing only minor discrepancies in its capacity to identify certain conditions. GPT-4o, another elite architecture, demonstrated an equitable diagnostic profile, yielding a precision of 0.95 and a recall of 0.88, which reinforces its trustworthiness for medical evaluation tasks. Similarly, GPT-3.5 exhibited robust diagnostic properties. While it did not match the sophistication of GPT-4, its competence in translating intricate symptom arrays into valid clinical assessments renders it a practical resource within medicine. GPT-3.5 aids clinical decision-making by offering dependable insights into medical scenarios, which can substantially support clinicians in interpreting patient presentations more effectively. Its stable performance underscores the reliability of well-tuned LLMs when processing clinical diagnostic tasks, highlighting the potential for artificial intelligence to act as a significant supportive tool in routine healthcare operations.

Conversely, while Gemini captured fewer correct diagnoses than the other models, it was characterized by its extraordinary diagnostic precision. This architecture operates under a conservative decision-making threshold, prioritizing output accuracy over sheer quantity. Such highly certain predictions make Gemini uniquely suited for critical clinical scenarios where error margins are virtually non-existent, such as diagnosing complex or rare syndromes where a false positive could result in severe clinical consequences. Gemini's methodology of minimizing false positives is highly valuable in medical settings where precision is vital and the threshold for error is minimal.

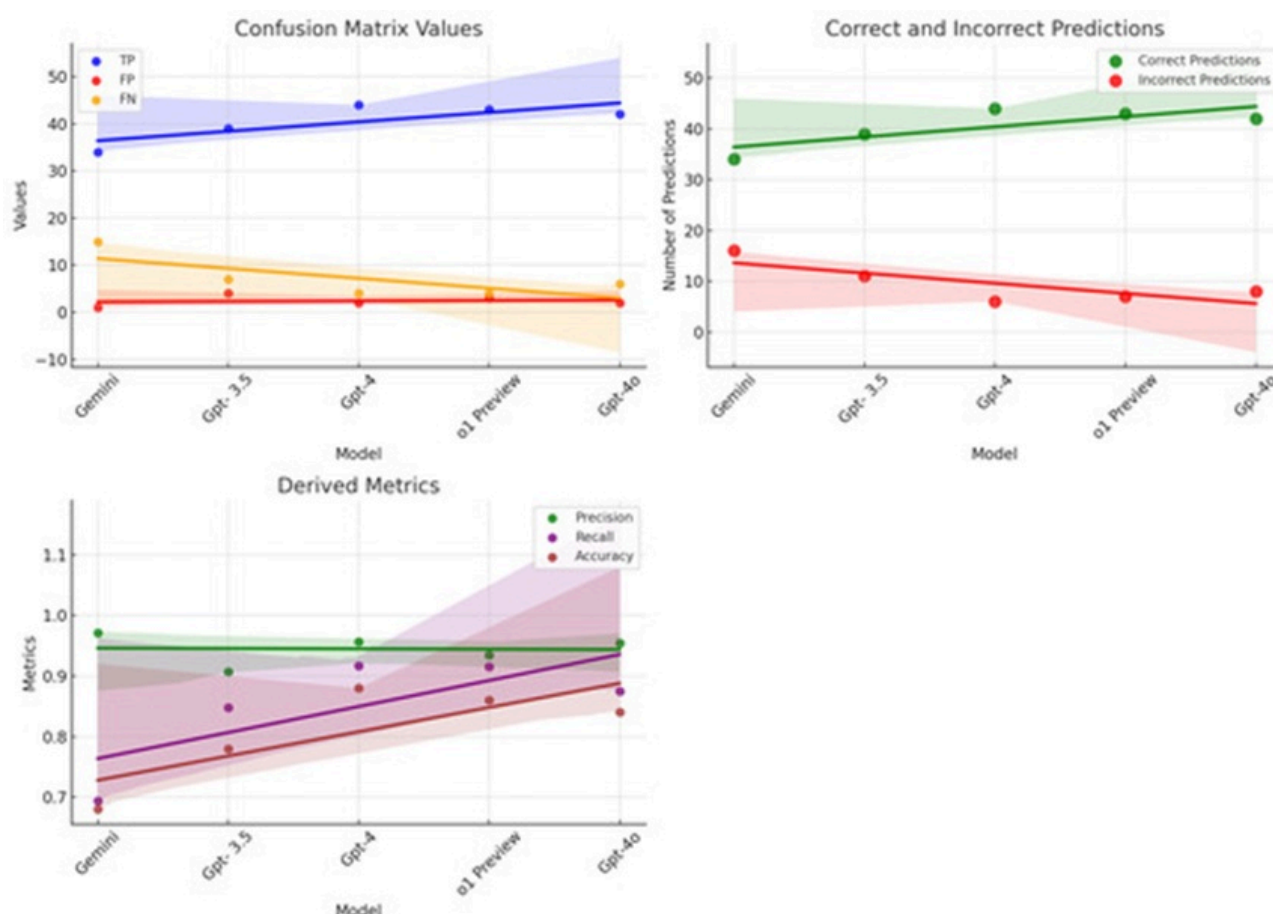


Figure 3. Performance metrics and predictions of models: scatter plots showcasing confusion matrix values, derived metrics, and the correctness of model predictions.

The aggregated performance data for these LLMs offers an encouraging outlook on the capacity of artificial intelligence to optimize clinical diagnostics. Incorporating these advanced architectures into healthcare systems could pave the way for faster, more precise, and highly dependable diagnostic pipelines. The broad interpretive reach of GPT-4, the steady utility of GPT-3.5, and the precise conservatism of Gemini together reflect the rapid evolution of artificial intelligence in medicine. The inclusion of o1 Preview and GPT-4o further broadens the operational flexibility of these models in clinical settings. The outcomes of this study not only highlight the major technological leaps AI has taken in deciphering the complexities of human health but also lay the groundwork for their integration into future medical practices. By elevating both the speed and exactness of clinical triage, these models can serve as indispensable assets for healthcare providers, ultimately driving better patient outcomes and reshaping contemporary healthcare delivery. As these technologies are continuously explored and optimized, their formal adoption into clinical environments promises to make medical care more effective, tailored, and universally accessible.

Comparative analysis

The comparative benchmarking of Gemini, GPT-3.5, GPT-4, o1 Preview, and GPT-4o throughout this research offers a clear overview of their diagnostic proficiencies, with each system revealing distinct operational advantages as quantified in the performance metrics detailed in **Table 2** and visually structured in **Figure 3**.

Table 2. Comparative performance of LLMs in digital diagnostics.

Model	Precision	Recall	F1 Score
Gemini	0.97	0.69	0.81
GPT-3.5	0.91	0.85	0.88
GPT-4	0.96	0.92	0.94
o1 Preview	0.93	0.91	0.92
GPT-4o	0.95	0.88	0.91

The exceptional volume of accurate classifications generated by GPT-4 serves as a strong indication of its exhaustive pre-training curriculum, which incorporated a highly diverse collection of clinical repositories. This rigorous computational preparation is directly demonstrated by its superior F1 score, which highlights the architecture's capacity to interpret intricate clinical terminology and reliably cross-reference symptom presentations with precise pathological states.

Figure 3 illustrates the comparative diagnostic accuracy across all evaluated models, highlighting the clear superiority of GPT-4 in generating accurate clinical classifications. This visualization highlights its exceptional capacity to unpack the intrinsic complexities that characterize the correlation between clinical symptoms and their underlying diagnoses. The o1 Preview model closely rivaled this performance, exhibiting elevated accuracy in pathological forecasting, while GPT-4o likewise demonstrated formidable capabilities across multiple diagnostic tasks.

Expanding upon this analytical review, **Figure 4** examines the distribution of certainty levels linked to the predictive outputs of each architecture. Within this framework, the certainty profiles for GPT-4, GPT-4o, o1 Preview, and GPT-3.5 were overwhelmingly concentrated within the “High” classification tier, illustrating their robust assertiveness and structural dependability when formulating diagnostic conclusions. Mirroring the performance of GPT-4, o1 Preview manifested distinct certainty in its predictive outputs, maintaining an flawless record within the “High” certainty tier with absolutely zero classifications falling into the low or medium ranges, which further emphasizes its dependability for high-stakes clinical tasks. Conversely, Gemini's strong preference for issuing exclusively high-certainty responses—notwithstanding a reduced total volume of generated predictions—underscored its unparalleled exactness. This operational characteristic is exceptionally vital within medical environments where the ramifications of a clinical miscalculation are severe, accentuating the clinical mandate for extreme precision and robust certainty during initial diagnostic assessments.

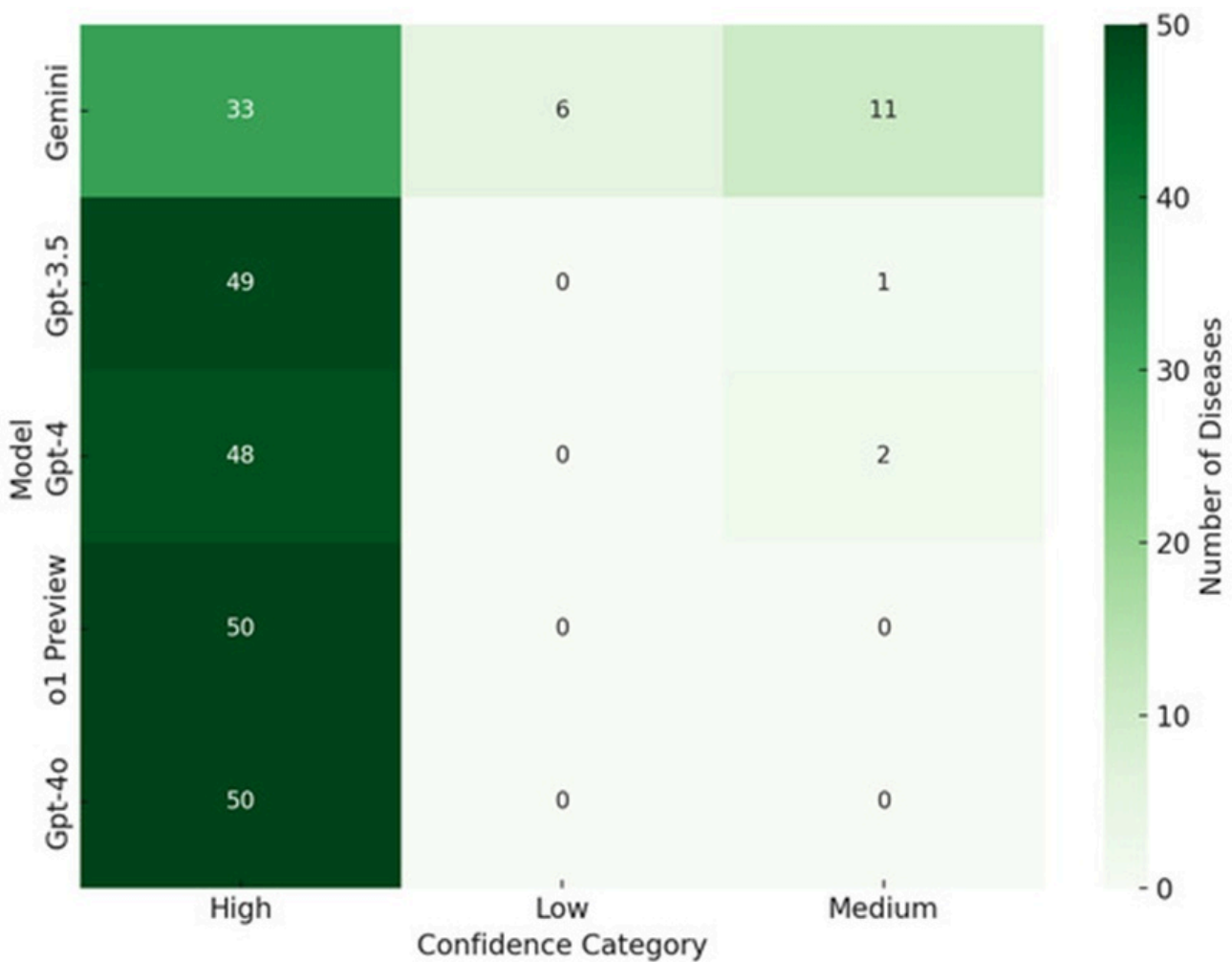


Figure 4. Model confidence category.

Nevertheless, Gemini's remarkable exactness came at the direct expense of its recall metrics, as corroborated by its statistical performance indices. This dynamic points to a highly conservative diagnostic methodology, wherein the model prioritizes absolute certainty over clinical breadth, which may inadvertently lead to omitting specific medical conditions during parsing. Meanwhile, GPT-3.5, though lacking the sophisticated architecture of GPT-4, demonstrated substantial practical utility in diagnostic processing by maintaining an equitable balance between precision and recall metrics. This stable performance establishes its functional value in environments where elite models like GPT-4 are either unavailable or structurally unnecessary.

Synthesizing these diagnostic observations, this investigation underscores the multifaceted clinical capabilities inherent across these diverse LLM architectures. GPT-4 distinguished itself as a highly versatile asset adaptable to varied medical fields, whereas the exceptional exactness of Gemini defines it as an indispensable utility for validating clinical determinations with maximum certainty. Simultaneously, the predictable and steady performance of GPT-3.5 ensures its sustained relevance within the rapidly shifting landscape of clinical artificial intelligence. The integration of o1 Preview and GPT-4o into this benchmarking matrix highlights the accelerating role of LLMs in delivering scalable, dependable decision support inside contemporary clinical environments.

The empirical data derived from this investigation supports the strategic, targeted deployment of LLMs within clinical infrastructures, specifically leveraging the distinct operational advantages of each respective model. Implementing such a tailored approach not only optimizes the baseline diagnostic pipeline but also elevates the broader standard of patient

care, driving the transition toward an AI-augmented medical ecosystem that prioritizes clinical exactness, operational efficiency, and patient safety.

Interpretation of results

This inquiry conducted a detailed evaluation of the functional proficiencies of five advanced Large Language Models (LLMs) tasked with identifying prevalent clinical pathologies from textual symptom arrays (Figure 5). The primary objective centered on gauging how effectively these computational systems could reinforce medical diagnostic workflows by interpreting symptom profiles and mapping them to corresponding disease entities. The premier model, GPT-4, clearly illustrated the transformative potential of artificial intelligence within clinical assessment through its advanced capacity to comprehend and process complex medical information arrays. The high efficiency of this model reflects its extensive pre-training across an broad spectrum of clinical scenarios, rendering it exceptionally proficient at aligning specific symptoms with their correct pathological designations. Following closely behind the top tier, o1 Preview exhibited remarkable diagnostic competencies, sustaining a powerful equilibrium between precision and recall that positions it as a highly reliable alternative for active clinical integration. GPT-4o also achieved elevated accuracy, displaying a well-balanced diagnostic profile that highlights its viability for real-time clinical applications. While operating with less technological sophistication than GPT-4, GPT-3.5 still manifested considerable capability in generating accurate medical assessments. Its capacity to deliver trustworthy determinations marks it as a practical supportive instrument in medical care, particularly within resource-limited areas where newer model generations may not yet be deployed. Finally, Gemini, characterized by its elite precision metrics, prioritized absolute accuracy within its predictive outputs—a vital trait in clinical contexts where diagnostic errors can cause severe patient complications.

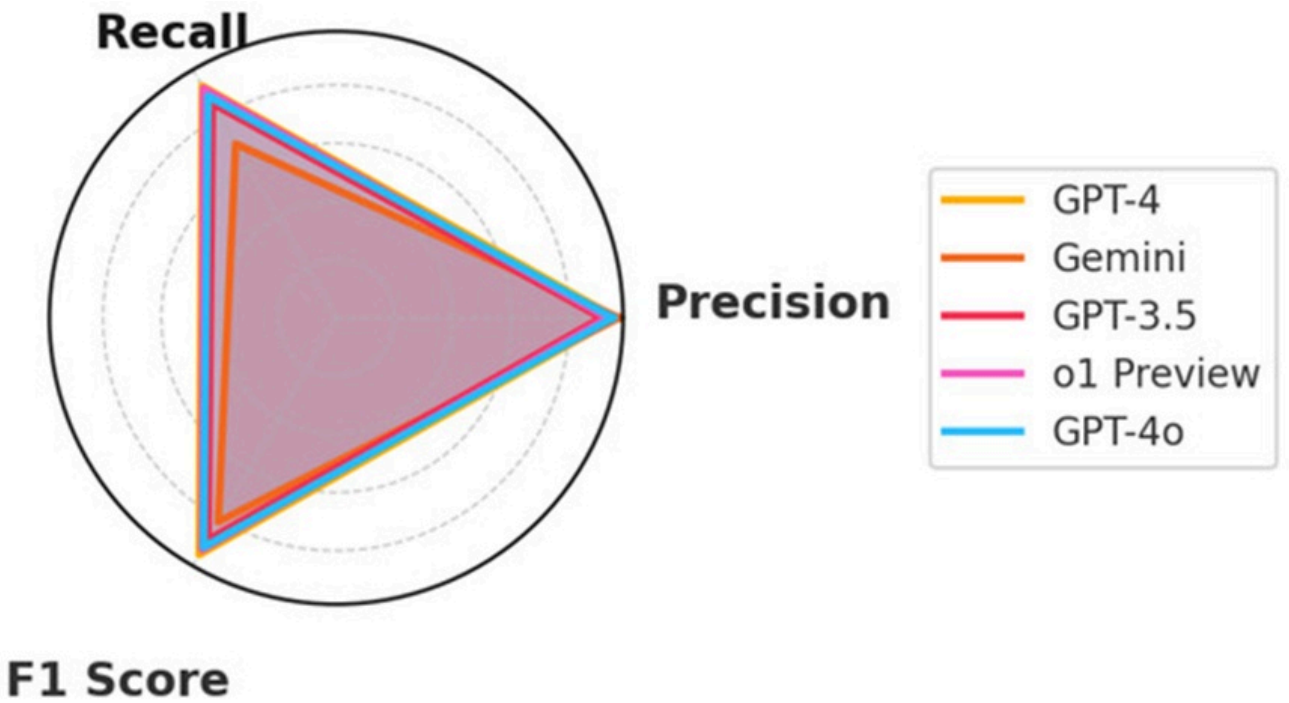


Figure 5. Performance metrics.

Enhancing diagnostic processes with large language models

The integration of Large Language Models (LLMs) into clinical workflows is poised to revolutionize healthcare delivery at the primary points of contact between patients and medical systems. Looking forward, LLMs will likely optimize both the

velocity and quality of initial medical triaging, facilitating rapid assessments of patient intake data and thereby substantially mitigating the administrative and cognitive burdens imposed on medical professionals, or enabling allied health staff to effectively manage triage roles. Within resource-constrained environments, particularly where prompt access to human clinical consultations is logistically unfeasible, LLMs could rapidly parse submitted symptom profiles to provide patients with initial indications regarding potential underlying pathologies [30, 43].

During the triage phase of these clinical presentations, LLMs could assist in organizing patient cohorts based on clinical urgency, effectively directing individuals to the appropriate tier of medical care. This optimization could directly translate into enhanced patient outcomes, including expedited clinical interventions, minimized appointment cancellations, and compressed waiting times prior to receiving active treatment. Enhanced patient education represents another major advantage of utilizing these conversational architectures. LLMs can be leveraged to supply patients with comprehensive information regarding their specific symptoms and potential diagnoses—additional knowledge that empowers individuals to more effectively understand and navigate their own health status [12].

Nevertheless, the systematic adoption of AI-driven instruments within clinical networks is inevitably accompanied by structural bottlenecks, requiring the careful navigation of several major dilemmas. For instance, compliance with established medical legislation, such as the Health Insurance Portability and Accountability Act (HIPAA) within the United States—a comprehensive federal statutory framework that dictates legal benchmarks and mandates the protection and secure transfer of protected health information (PHI) across medical systems—remains paramount [14]. Every piece of data harvested by LLMs during clinical dialogues must strictly conform to HIPAA standards; these systems must process vital medical insights and patient interactions under conditions of total data security and absolute confidentiality.

Furthermore, while LLMs offer valuable analytical support, their incorporation must be structured to augment rather than replace the human dimensions of care delivery, sharpening physician decision-making and diagnostic intuition rather than displacing the clinician. The error margins of LLMs when confronted with highly complex clinical presentations must be continuously cross-referenced and validated against established gold-standard medical benchmarks [28]. Any significant deviation from accepted clinical standards introduces the immediate risk of a misdiagnosis. Consequently, these models require continuous maintenance and updating through cyclical auditing and anomaly flagging.

Sustaining a dedicated, cross-disciplinary alliance among software engineers, practicing clinicians, and regulatory authorities is equally critical for the successful deployment of LLMs in medicine. Standardized, interdisciplinary workflows are vital for designing AI utilities that are structurally stable and clinically functional, as well as for executing appropriate validation metrics and guaranteeing ethical application. Rigorously testing each architecture must extend far beyond basic optimization of technical or computational benchmarks to confirm that the tool can genuinely and safely assist medical practitioners within authentic clinical workflows [44].

In conclusion, embedding LLMs within healthcare frameworks represents a highly promising pathway for optimizing diagnostic exactness and refining patient management. However, successfully overcoming the inherent challenges of this technology requires meticulous operational planning and the establishment of robust validation protocols. Adhering to these structured processes will maximize the clinical utility of artificial intelligence in medicine, ultimately elevating outcomes for patients and healthcare systems alike.

Limitations of the study

A primary constraint of this investigation centers on the selection of relatively straightforward, low-acuity pathologies for the model evaluation matrix, an approach that fails to capture the intricate clinical dimensions of diagnosing severe or long-term chronic conditions. The diagnosis of chronic diseases typically demands the interpretation of convoluted, overlapping symptom arrays that cannot be safely unraveled without a holistic review of a patient's longitudinal medical records and supplementary objective diagnostics. Furthermore, this study's exclusive focus on linguistic text processing overlooks the inherently multimodal framework of traditional medical diagnostics, which routinely fuses verbal

symptomatology with physical assessments, detailed personal case files, and visual clinical imaging. Consequently, future iterations of medical artificial intelligence must prioritize the incorporation of these heterogeneous information channels to achieve more precise and comprehensive diagnostic outputs.

Future research directions

Subsequent scholarly inquiries should prioritize the architecture of LLMs capable of cross-referencing diverse healthcare data modalities beyond standalone textual inputs. Crucial to this evolution will be the systematic assimilation of visual data from radiological and diagnostic imaging to construct highly integrated diagnostic ecosystems. Additionally, engineering AI frameworks with the capacity to interpret and process data across diverse linguistic structures and heterogeneous cultural frameworks represents a vital milestone for achieving universal applicability. Because the current study relied on a baseline proof-of-concept repository, our upcoming research agenda involves deploying these benchmarks against a broader, larger-scale dataset to better judge algorithmic functionality across highly complex pathological states. Broadening the informational data pool will yield granular insights regarding the specific analytical strengths and architectural constraints of LLMs when confronted with a wider spectrum of clinical presentation.

Furthermore, it remains imperative that the development of these computational architectures strictly conforms to rigorous ethical guidelines, with an absolute emphasis on safeguarding patient data security. Striking an optimal balance between the deployment of healthcare AI and the strict maintenance of patient confidentiality constitutes a vital area of upcoming research. Lastly, embedding these advanced language models directly into active clinical settings to systematically measure their performance metrics, impact on institutional throughput, and patient care outcomes will generate crucial empirical data regarding their real-world utility and operational limitations.

This investigation highlights both the profound opportunities and the significant roadblocks associated with introducing advanced AI frameworks into the diagnostic space. Driven by methodical software development and an unyielding commitment to ethical protocols, these digital instruments could substantially upgrade clinical efficiency and diagnostic precision. Ongoing exploration and iterative refinement of these underlying models are likely to facilitate their successful adoption into routine clinical practice, yielding measurable advantages for healthcare systems and patients alike.

Conclusion

This detailed evaluation into the behavioral and linguistic dynamics of Large Language Models (LLMs)—specifically Gemini, GPT-3.5, GPT-4, o1 Preview, and GPT-4o—has yielded an all-encompassing perspective on their capacity to optimize digital diagnostic frameworks within modern healthcare. Through a rigorous assessment matrix, this study has shed light on the relative proficiencies of these models in isolating common pathological symptoms while simultaneously clarifying their core operational boundaries. The definitive findings indicate that architectures like GPT-4 hold considerable potential for parsing medical terminology with elevated accuracy, representing a major milestone toward the delivery of immediate, accessible clinical orientation. Simultaneously, both o1 Preview and GPT-4o delivered impressive benchmarks, demonstrating powerful diagnostic competence characterized by an equitable distribution of precision and recall metrics, thereby validating their suitability for real-time clinical workflows. The exceptional exactness exhibited by the specialized Gemini model suggests that engineering niche, domain-restricted LLMs is a highly viable strategy for supplying highly reliable decision-support utilities. Meanwhile, GPT-3.5, though naturally eclipsed by the technical benchmarks of subsequent model generations, still displayed respectable diagnostic capabilities that reflect the rapid evolutionary velocity of medical artificial intelligence. Collectively, these empirical insights reinforce the disruptive potential of LLMs in digital triage, demonstrating their capacity to act as a valuable adjunct to traditional diagnostic protocols while broadening the availability and quality of patient care.

Nevertheless, transitioning from theoretical promise to active clinical implementation necessitates overcoming formidable technical and regulatory barriers, including the integration of multimodal data channels, the resolution of ethical dilemmas, and strict alignment with stringent healthcare governance protocols. Because this investigation was limited by its reliance

on textual data streams and manual validation processes, it establishes a transparent directive for subsequent research endeavors. The next phase of development must expand the interpretive capabilities of LLMs far beyond textual domains, unifying visual and empirical data channels to closely mirror the multi-faceted reality of clinical medicine. Moreover, this research emphasizes the critical necessity of building ethically grounded, culturally responsive, and polyglot LLM frameworks to effectively serve global patient cohorts. Within the broader movement to harness artificial intelligence for clinical purposes, LLMs must not be viewed as autonomous diagnostic solutions, but rather as collaborative instruments engineered to augment the baseline expertise of medical professionals. As computational medicine advances, embedding these models within active clinical workflows will be essential to fully quantify their day-to-day dependability and functional value. Sustained by iterative optimization and conscientious deployment, LLMs are uniquely positioned to occupy a central role in redefining future healthcare delivery, rendering diagnostics increasingly accessible, highly accurate, and thoroughly patient-centric.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
GPT	Generative Pre-trained Transformer
HIPAA	Health Insurance Portability and Accountability Act
NLP	Natural Language Processing
AI	Artificial Intelligence
CDC	Centers for Disease Control and Prevention
WHO	World Health Organization
EHR	Electronic Health Records
FP	False positive
TP	True positive
FN	False negative

Acknowledgements

None

Conflict of interest

Author Abul Ehtesham was employed by the Davey Tree Expert Company. 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.

Financial support

None

Ethics statement

None

Received: 08 Apr 2022

Revised: 04 Jun 2022

Accepted: 12 Jul 2022

Published online: 15 September 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey. arXiv [Preprint]. 2024. arXiv:2402.06196.
- Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv [Preprint]. 2023. arXiv:2303.18223.
- Singh A. Exploring language models: a comprehensive survey and analysis. In: Proceedings of the 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE); 2023; Chennai, India. p.1-4.
- Hadi MU, Al Tashi Q, Qureshi R, Shah A, Muneer A, Irfan M, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. TechRxiv [Preprint]. 2024. <https://doi.org/10.36227/techrxiv.23589741>.
- OpenAI. GPT-4 technical report. arXiv [Preprint]. 2023. arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>.
- Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res. 2024;26:e56764. <https://doi.org/10.2196/56764>.
- Gemini Team. Gemini: a family of highly capable multimodal models. arXiv [Preprint]. 2023. arXiv:2312.11805.
- Webster P. Six ways large language models are changing healthcare. Nat Med. 2023;29(12):2969-71. <https://doi.org/10.1038/s41591-023-02700-1>.
- Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. Sci Rep. 2024;14(1):2840. <https://doi.org/10.1038/s41598-024-52853-3>.
- Cui H, Fang X, Xu R, Kan X, Ho JC, Yang C. Multimodal fusion of EHR in structures and semantics: integrating clinical records and notes with hypergraph and LLM. arXiv [Preprint]. 2024. arXiv:2403.08818.
- Montagna S, Ferretti S, Klopfenstein LC, Flori A, Pengo MF. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. In: Proceedings of the 2023 ACM Conference on Information Technology for Social Good; 2023; Lisbon,

Portugal. p.205-12.

Kusa W, Mosca E, Lipani A. "Dr LLM, what do I have?": the impact of user beliefs and prompt formulation on health diagnoses. In: Khosla S, editor. *Proceedings of the Third Workshop on NLP for Medical Conversations*; 2023 Nov; Bali, Indonesia. Association for Computational Linguistics; 2023. p.13-9.

Lai T, Shi Y, Du Z, Wu J, Fu K, Dou Y, et al. Psy-LLM: scaling up global mental health psychological services with AI-based large language models. *arXiv [Preprint]*. 2023. arXiv:2307.11991.

Humphrey BA. Data privacy vs. innovation: a quantitative analysis of artificial intelligence in healthcare and its impact on HIPAA regarding the privacy and security of protected health information [dissertation]. Moon Township (PA): Robert Morris University; 2021.

Dhakal U, Singh AK, Devkota S, Sapkota Y, Lamichhane B, Paudyal S, et al. GPT-4's assessment of its performance in a USMLE-based case study. *arXiv [Preprint]*. 2024. arXiv:2402.09654.

Denecke K, May R, LLMHealthGroup, Rivera Romero O. Potential of large language models in health care: Delphi study. *J Med Internet Res*. 2024;26:e52399.
<https://doi.org/10.2196/52399>.

Meskó B. The role of artificial intelligence in precision medicine. *Expert Rev Precis Med Drug Dev*. 2017;2(5):239-41.
<https://doi.org/10.1080/23808993.2017.1380516>.

Singh A, Ehtesham A, Mahmud S, Kim JH. Revolutionizing mental health care through LangChain: a journey with a large language model. In: *Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*; 2024; Las Vegas, NV, USA. IEEE; 2024. p.73-8.
<https://doi.org/10.1109/CCWC60891.2024.10427865>.

de Curtò J, de Zarzà I, Roig G, Cano JC, Manzoni P, Calafate CT. LLM-informed multi-armed bandit strategies for non-stationary environments. *Electronics*. 2023;12:2814.

Chiu YY, Sharma A, Lin IW, Althoff T. A computational framework for behavioral assessment of LLM therapists. *arXiv [Preprint]*. 2024. arXiv:2401.00820.

Batsis JA, Mackenzie TA, Emeny RT, Lopez-Jimenez F, Bartels SJ. Low lean mass with and without obesity, and mortality: results from the 1999-2004 National Health and Nutrition Examination Survey. *J Gerontol A Biol Sci Med Sci*. 2017;72(10):1445-51.

Baharudin N, Mohamed-Yassin MS, Daher AM, Ramli AS, Khan NMN, Abdul-Razak S. Prevalence and factors associated with lipid-lowering medications use for primary and secondary prevention of cardiovascular diseases among Malaysians: the REDISCOVER study. *BMC Public Health*. 2022;22:228.

Abbasian M, Azimi I, Rahmani AM, Jain R. Conversational health agents: a personalized LLM-powered agent framework. *arXiv [Preprint]*. 2024. arXiv:2310.02374.

Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. *iScience*. 2024;27(5):109713.
<https://doi.org/10.1016/j.isci.2024.109713>.

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3:141.
<https://doi.org/10.1038/s43856-023-00370-1>.

Reese JT, Danis D, Caufield JH, Groza T, Casiraghi E, Valentini G, et al. On the limitations of large language models in clinical diagnosis. *medRxiv [Preprint]*. 2023.

Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int J Med Inform.* 2020;141:104431.

Singh AK, Lamichhane B, Devkota S, Dhakal U, Dhakal C. Do large language models show human-like biases? Exploring confidence-competence gap in AI. *Information.* 2024;15:92.

Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology-a recent scoping review. *Diagn Pathol.* 2024;19:43.

Jo E, Epstein DA, Jung H, Kim YH. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*; 2023; Hamburg, Germany. ACM; 2023. p.1-16.

Baric-Parker J, Anderson E. Patient data-sharing for AI: ethical challenges, Catholic solutions. *Linacre Q.* 2020;87:471-81.

Jin M, Yu Q, Shu D, Zhang C, Fan L, Hua W, et al. Health-LLM: personalized retrieval-augmented disease prediction system. *arXiv [Preprint].* 2024. arXiv:2402.00746.

Kim Y, Xu X, McDuff D, Breazeal C, Park HW. Health-LLM: large language models for health prediction via wearable sensor data. *arXiv [Preprint].* 2024. arXiv:2401.06866.

Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler JA, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2024;8(1):31.
<https://doi.org/10.1145/3643540>.

Ghosh A, Acharya A, Jain R, Saha S, Chadha A, Sinha S. CLIPSyntel: CLIP and LLM synergy for multimodal question summarization in healthcare. *arXiv [Preprint].* 2023. arXiv:2312.11541.

Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Sci Rep.* 2024;14:9330.

Yuan J, Tang R, Jiang X, Hu X. Large language models for healthcare data augmentation: an example on patient-trial matching. *AMIA Annu Symp Proc.* 2024;2023:1324-33.

Centers for Disease Control and Prevention. Symptoms of diseases [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; [cited 2024 Oct 9]. Available from: CDC website.

World Health Organization. Disease symptoms and information [Internet]. Geneva: World Health Organization; [cited 2024 Oct 9]. Available from: WHO website.

Mayo Clinic. Symptoms of common diseases [Internet]. Rochester (MN): Mayo Foundation for Medical Education and Research; [cited 2024 Oct 9]. Available from: Mayo Clinic website.

Cleveland Clinic. Disease symptoms and conditions [Internet]. Cleveland (OH): Cleveland Clinic; [cited 2024 Oct 9]. Available from: Cleveland Clinic website.

Johns Hopkins Medicine. Symptoms and disease information [Internet]. Baltimore (MD): Johns Hopkins Medicine; [cited 2024 Oct 9]. Available from: Johns Hopkins Medicine website.

Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11(6):887.

Frantzidis CA, Bamidis PD. Description and future trends of ICT solutions offered towards independent living: the case of LLM project. In: Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments; 2009; Corfu, Greece. p.59.