Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Occam's Razor in AI-assisted complex diagnosis: a comparative effectiveness study of single large language models versus multi-agent systems in resource-constrained primary care settings

Tengfei Cai
Weifang People's Hospital
Naiguang Zhang
Weifang People's Hospital
Yansheng Li
iMEDWAY Technology Co.,Ltd
Yanmin Li
Weifang People's Hospital
Xiaoyan Li
Weifang People's Hospital
Bo Liu
wfsrmyy@outlook.com

Weifang People's Hospital

Research Article

Keywords: Occam's Razor, AI, diagnosis, large language model, multi-agent system, primary care settings

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10586813/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Primary care physicians in resource-constrained settings, particularly within low-income and middle-income countries (LMICs), frequently encounter a "diagnostic gap" when managing complex, rare, or multisystemic pathologies. While Large Language Models (LLMs) demonstrate significant potential to augment clinical reasoning, current state-of-the-art solutions rely predominantly on high-bandwidth cloud infrastructure, limiting their deployment in regions with unstable internet connectivity and strict data sovereignty regulations. The prevailing technological consensus in computer science suggests that "Agentic Workflows" or Multi-Agent Systems (MAS)—which orchestrate multiple models to simulate collective reasoning—inherently offer superior accuracy and safety compared to single models.

Objectives

This study aimed to test the hypothesis that complex multi-agent systems would outperform single localised models in diagnostic accuracy within offline, hardware-limited environments. We also sought to evaluate the comparative efficacy, safety, and cost-effectiveness of single LLMs versus MAS in resource-constrained primary care settings.

Methods

We conducted a prospective comparative benchmarking study using the DiagnosisArena dataset, comprising 915 complex clinical cases across 28 medical specialties. To simulate a secure, offline primary care environment, we evaluated five locally deployed single open-source LLMs (GPT-oss-20b, Llama3.1-70B, Qwen3-32B, DeepSeek-R1-32B, Gemma3-27B) against two Multi-Agent architectures: a Standard voting ensemble and a novel hierarchical Adaptive Weighted System. All models were hosted on a local server (4×NVIDIA A100) using the Dify platform. Performance was adjudicated against a Reference Standard established by the consensus of three board-certified physicians using a dual-metric system: a 10-point Diagnostic Recall Scale and a comprehensive Hallucination/Safety Index. Inference latency and computational resource utilisation were recorded to assess cost-effectiveness. Statistical analyses employed one-way ANOVA with post-hoc pairwise comparisons using the Least Significant Difference method and paired t-tests.

Results

Contrary to the hypothesis that architectural complexity yields diagnostic precision, single high-performance models significantly outperformed complex ensembles. The single GPT-oss-20b model achieved the highest Diagnostic Recall Score (mean 4.68 [SD 3.82]), statistically surpassing the Adaptive

Weighted Multi-Agent System (4.13 [SD 3.43]; $p < 0.001$) and smaller models such as Gemma3-27B (2.89 [SD 3.89]; $p < 0.001$). The Adaptive System, despite utilising dynamic routing, failed to outperform the median score of human physicians (4.22 [SD 3.62]; $p = 0.432$). Furthermore, the inclusion of mid-tier models in the adaptive workflow introduced an "ensemble degradation" effect, significantly lowering the Safety Score compared to the single GPT-oss-20b model (4.99 vs 5.50; $p < 0.001$) and reducing the rate of Top-1 correct diagnoses from 51.58% to 46.89%. Crucially, the single GPT-oss-20b model demonstrated superior efficiency with an average inference time of 30 seconds per case, compared to 200 seconds for the Standard Multi-Agent System—representing an 85% reduction in latency.

Conclusions

In the context of clinical diagnosis, architectural complexity does not equate to clinical utility. We identified a phenomenon of "ensemble degradation," where integrating mid-tier models into ensembles dilutes the reasoning capabilities of strong base models through the introduction of diagnostic noise. For global health equity, implementation strategies should prioritise "Lean AI"—localising a single, robust open-source model—rather than orchestrating computationally expensive agent swarms. This approach provides a safer, more accurate, and scientifically validated path for bridging the diagnostic gap in resource-constrained primary care.

Introduction

The diagnostic odyssey for patients presenting with complex, rare, or multisystemic pathologies often begins—and unfortunately stalls—at the primary care level [1]. General practitioners (GPs) form the operational backbone of healthcare systems globally; however, they face a persistent structural paradox [2]. They are the mandated first point of contact for undifferentiated symptoms, yet they often lack the specialised resources and diagnostic tools required for atypical conditions [3]. This "diagnostic gap" is a significant driver of global health inequality, particularly in Low- and Middle-Income Countries (LMICs) where specialist referral pathways are fragmented [4, 5].

The rapid evolution of Large Language Models (LLMs) offers a potentially transformative solution to bridge this gap [6, 7]. Studies have demonstrated that LLMs can rival human specialists in clinical reasoning [8]. However, the majority of state-of-the-art models are deployed via cloud APIs [9]. This dependence presents critical challenges for grassroots deployment: patient data privacy regulations, prohibitive costs, and reliance on stable internet connectivity are major barriers in rural clinics of the Global South [10, 11]. Consequently, the development of localised, offline-capable AI solutions has become an urgent priority [12].

In the pursuit of higher diagnostic accuracy, the field has increasingly turned towards architectural complexity, specifically Multi-Agent Systems (MAS) that simulate a medical board [13, 14]. The theoretical premise is based on "collective intelligence"—that a diverse ensemble of models will cancel

out individual errors and converge on the truth [15]. This suggests that to achieve expert-level performance, systems must move beyond single-model prompting to complex agent orchestration.

However, this assumption remains untested in clinical environments constrained by consumer-grade hardware and limited computational budgets [16]. A critical question persists: Does the orchestration of multiple agents actually enhance diagnostic signal, or does it introduce noise, latency, and unnecessary computational debt? This study aimed to test the hypothesis that complex multi-agent systems would outperform single localised models in diagnostic accuracy. We conducted a rigorous comparative effectiveness study to determine whether the complexity of multi-agent orchestration is a clinical necessity or a technological redundancy.

Methods

Study Design and Data Sources

We employed a prospective, multi-arm comparative experimental design. The study utilised the DiagnosisArena dataset, a comprehensive validated corpus comprising 915 complex clinical cases sourced from peer-reviewed case reports in high-impact medical journals (e.g., The New England Journal of Medicine, The Lancet), covering 28 medical specialties [17].

The dataset is characterised by high clinical heterogeneity, covering 28 medical specialties including Internal Medicine, General Surgery, Neurology, Cardiovascular Medicine, and Infectious Diseases. Each case typically involves multisystemic involvement, non-specific initial symptoms, and a requirement for multi-step deductive reasoning.

To establish a rigorous ground truth, a Reference Standard (Gold Standard) was annotated for each case, consisting of the final histopathologically or clinically confirmed diagnosis and key diagnostic evidence. Annotation was performed independently by three attending physicians (or higher rank) with a minimum of five years of clinical experience. Discrepancies were adjudicated by a fourth senior physician (Associate Chief Physician rank) to ensure 100% consensus on the correct diagnosis and the critical evidence required to reach it.

Ethical Considerations:This study involved the benchmarking of AI algorithms on de-identified, retrospective case data publicly available or synthesised from medical literature. No human subjects were prospectively enrolled, and no clinical interventions were performed on patients. As such, the study was exempt from institutional review board (IRB) approval for human subject research, in accordance with the Declaration of Helsinki.

AI Model Architectures and Local Deployment

To simulate a secure, resource-constrained primary care environment, all models were deployed locally using the Dify platform (Community Edition v0.6.0) on a standalone server equipped with 4×NVIDIA A100 GPUs. This setup was chosen to mirror a "hub-and-spoke" deployment model where a regional hospital server might support multiple offline clinics. Crucially, this ensured strict data privacy (no data left the local server) and eliminated dependence on external cloud connectivity.

Inference parameters were standardised across all models to ensure fair comparison: temperature = 0.2 (to minimise hallucination while allowing creative reasoning), top_p = 0.9, top_k = 50, and max_tokens = 1024. The prompt structure utilised a zero-shot "Chain-of-Thought" (CoT) framework, instructing the models to "think step-by-step" before concluding with a differential diagnosis.

We evaluated seven distinct experimental configurations:

1. Single Model Group (n = 5):DeepSeek-R1-32B, Qwen3-32B, Llama3.1-70B, Gemma3-27B, and GPT-oss-20b. These open-source models were selected to represent a range of parameter sizes (27B to 70B) and architectures, providing a gradient of computational capability.
2. Multi-Agent System (MAS) Group (n = 2):

Standard Multi-Agent System:An ensemble approach aggregating outputs from the top-performing single models via uniform voting rules. This simulates a "Medical Board" where every member has an equal vote.

Adaptive Weighted Multi-Agent System:A specialised architecture utilising dynamic routing and performance-based weighting.

The Adaptive Weighting Algorithm

The Adaptive Weighted Multi-Agent system employed a hierarchical decision architecture designed to test the hypothesis that "smart routing" could optimise performance by assigning cases to the most suitable models. A gating classifier (based on a lightweight GPT-oss-20b instance) first categorised incoming cases into one of three clinical workflows: Internal Medicine, Surgery, or Other/General.

Within each workflow, model contributions were governed by a "Model Weight Adaptive" mechanism. The weights were calculated based on a formula derived from a preliminary pilot study of 50 training cases (separate from the test set):

W=Scoremodel/∑Scoreall (1)

Where the Base Score is calculated as:

Scoremodel = R×0.6 + (1-H/100)×0.4 (2)

Where R represents the Recall Rate (percentage of key clinical cues identified) and H represents the Hallucination Rate. The pilot study indicated that a 60/40 weighting balance between accuracy and safety provided the most stable performance during calibration. For instance, in the Surgery workflow, the stronger Llama3.1-70B was assigned a weight of 0.58, while the mid-tier Qwen3-32B was assigned 0.42. This weighting was applied to the probability logits during the final consensus generation.

Evaluation Strategy and Metrics

We developed a rigorous dual-metric evaluation system to assess both clinical efficacy and patient safety:

1. Diagnostic Recall Rate Scale (10-point scale): This primary efficacy endpoint assessed the alignment between model output and the Reference Standard.

Final Diagnosis Recall (0–6 points): Scored based on the rank of the correct diagnosis within the model's Top-5 output. An exact Top-1 match yielded 6 points; Top 2–3 yielded 4 points; Top 4–5 yielded 2 points; failure to list yielded 0 points.

Key Evidence Recall (0–4 points): Calculated based on the percentage of Key Diagnostic Clues identified in the model's reasoning chain. This metric ensures the model arrived at the right answer for the right reasons, penalising "lucky guesses."

2. Hallucination and Safety Scale: A composite score (0–10) incorporating two dimensions:

Ranking Level (Level A–D): Level A indicates the correct diagnosis is Top 1; Level D indicates it is missing from the Top 5.

Severity Level (0–5):A risk-stratification of errors. We defined "Level 5" as Dangerous Errors (e.g., suggesting thrombolysis for a haemorrhagic stroke). Such critical errors triggered a severe penalty, capping the maximum safety score at 2 points regardless of diagnostic ranking.

3. Cost-Effectiveness Metrics: To quantify the operational burden, we recorded Inference Latency (average seconds per case) and Peak GPU Memory Utilisation. These metrics serve as proxies for the hardware capital expenditure (CAPEX) and energy operational expenditure (OPEX) required for deployment.

Scoring was automated using the DeepSeek V3.2 API via a standardised evaluation prompt template. To validate this automated evaluation, a stratified random sample of 10% of cases (n = 92) underwent manual re-evaluation by human experts. The agreement rate (Kappa statistic) between the automated scorer and human experts exceeded 0.90, validating the reliability of the automated pipeline.

Statistical Analysis

Data were analysed using Python 3.12. Continuous variables (Recall Scores, Safety Scores, Latency) were expressed as mean ± standard deviation (SD). Group comparisons were performed using one-way analysis of variance (ANOVA). Post-hoc pairwise comparisons utilised the Least Significant Difference (LSD) method for multiple groups and paired t-tests for direct model-to-model comparisons. Categorical variables (e.g., proportion of Level A cases) were analysed using the χ^2test. A two-sided p-value of < 0.05 was considered statistically significant.

Results

Diagnostic Accuracy and Recall Performance

The analysis of 915 complex cases revealed a distinct performance hierarchy that correlated more strongly with individual model capability than with architectural complexity. The single-agent GPT-oss-20b model achieved the highest overall performance, yielding a mean Total Recall Score of 4.68 (SD 3.82) out of 10. This was closely followed by Llama3.1-70B (4.62 ± 3.79). Statistical analysis using paired t-tests revealed no significant difference between these two top-performing single models ($p = 0.337$). Notably, both single models outperformed the human physician cohort, which achieved a mean score of 4.22 (SD 3.62), although the difference for GPT-oss-20b vs. Physicians bordered on significance.

Conversely, the smaller parameter models demonstrated significantly inferior performance. DeepSeek-R1-32B and Gemma3-27B achieved Total Recall Scores of 2.88 (SD 3.33) and 2.89 (SD 3.89), respectively. Both models scored significantly lower than the GPT-oss-20b baseline ($p < 0.001$ for both comparisons).

The distribution of scores further elucidates this disparity. GPT-oss-20b placed 28.42% (n = 260) of cases in the "High Score Zone" (9–10 points), indicating a flawless diagnosis and reasoning chain. In stark contrast, Gemma3-27B only achieved this in 16.83% (n = 154) of cases, with the majority (65.46%) falling into the "Low Score Zone" (0–2 points), representing a failure to diagnose. This validates the premise that model parameter size is a critical determinant of reasoning capability in complex medical scenarios.

Comparative Analysis: The Failure of Complexity

A central and counter-intuitive finding of this study is that architectural complexity did not correlate with diagnostic utility.

The Standard Multi-Agent System—which aggregated votes from multiple models—achieved a Total Recall Score of 4.54 (SD 3.65). When compared to the single GPT-oss-20b model, this difference was not statistically significant ($p = 0.249$), indicating that the computational overhead of the multi-agent ensemble provided no measurable gain in diagnostic accuracy.

Furthermore, the Adaptive Weighted Multi-Agent System—which theoretically should have optimised performance by routing cases—resulted in statistically significant performance degradation. The Adaptive system achieved a Total Recall Score of 4.13 (SD 3.43), which was significantly lower than the single GPT-oss-20b model ($p < 0.001$). Paired t-tests indicated that the Adaptive System was statistically indistinguishable from the median performance of human physicians ($p = 0.432$).

Detailed analysis of the weighting mechanism reveals the cause: by incorporating inputs from mid-tier models (e.g., Qwen3-32B) into the decision matrix, even with lower weights, the system effectively diluted the correct signals generated by the stronger models. For example, in surgical cases, while Llama3.1-70B might correctly identify a rare complication, the concurrent incorrect outputs from Qwen3-32B and DeepSeek-R1-32B pulled the weighted consensus away from the correct diagnosis.

Hallucination Rates and Safety Profile

Safety is the paramount ethical obligation in clinical AI. Our analysis of the Hallucination and Safety Scale mirrored the trends observed in diagnostic accuracy.

The single GPT-oss-20b model achieved a Comprehensive Safety Score of 5.50 (SD 4.58). In terms of "Level A" accuracy (where the Reference Standard is the unequivocal Top-1 output), GPT-oss-20b achieved 51.58% (n = 472).

The Adaptive Weighted System showed a concerning decline in Top-1 accuracy to 46.89% (n = 429), and its Comprehensive Safety Score (4.99 ± 4.28) was significantly lower than that of the single GPT-oss-20b model ($p < 0.001$).

The lowest performing model, Gemma3-27B, exhibited a "Level D" rate (failure to list the correct diagnosis in the Top-5) of 56.61%, significantly higher than the 36.83% Level D rate observed in the GPT-oss-20b group. These data confirm that while single high-performance models provide a robust safety profile, the introduction of adaptive complexity involving mid-tier models introduces diagnostic noise rather than correcting it.

Cost-Effectiveness and Efficiency Analysis

In resource-constrained settings, computational efficiency is not merely a technical metric but a determinant of feasibility. We observed a stark contrast in inference latency.

Single Models: GPT-oss-20b and Llama3.1-70B demonstrated high efficiency, with average inference times of 30 seconds and 43 seconds per case, respectively.

Multi-Agent Systems: The Standard Multi-Agent System required an average of 200 seconds per case—nearly 7 times the latency of GPT-oss-20b. The Adaptive System averaged 104 seconds.

From a hardware perspective, the Multi-Agent systems required the simultaneous loading of multiple models into VRAM, necessitating the full capacity of the 4×A100 server (approx. 320GB VRAM). In

contrast, the single GPT-oss-20b model could theoretically run on a dual consumer-grade GPU setup (e.g., 2×RTX 3090/4090, approx. 48GB VRAM) via quantisation techniques. This represents a potential hardware cost reduction of approximately 80–90%, making the "Single Model" approach significantly more viable for district hospitals in LMICs.

Table 1 summarizes the diagnostic performance, safety profile, inference latency and efficiency classification of all included large language models, as well as the benchmark data from human physicians.

Efficiency Status is a qualitative assessment derived from the balance between diagnostic performance (Recall Score) and computational resource utilization (Inference Latency)

Discussion

The Fallacy of Complexity in Clinical AI

Our findings present a significant challenge to the prevailing technological orthodoxy in medical artificial intelligence. While the dominant narrative in computer science advocates for "Agentic Workflows," "Collective Intelligence," and "Mixture of Experts" as the path to Artificial General Intelligence (AGI), our results demonstrate that in the specific domain of complex clinical diagnosis, architectural complexity does not equate to clinical utility. Specifically, the Adaptive Weighted Multi-Agent System did not outperform the single best-in-class model; in fact, it statistically degraded diagnostic performance (*p* <0.001 vs GPT-oss-20b).

Recent studies have reported promising results with multi-agent frameworks in medical tasks [18, 19]; however, these systems typically employ homogeneous high-capacity agents (e.g., all GPT-4 based), whereas our adaptive architecture incorporated heterogeneous models of varying capability. The divergence between our findings and those reporting benefits of MAS may be attributed to this distinction: prior successes often relied on identical strong base models, while our configuration mixed heterogeneous models—a design that recent work suggests is prone to "negative ensemble transfer" or diagnostic signal dilution when weaker models are introduced into the voting matrix [23, 24].

This phenomenon can be best described as "Ensemble Degradation" or the "Dilution Effect." In our adaptive architecture, high-performing models (e.g., Llama 3.1-70B) were computationally forced to collaborate with mid-tier models (e.g., Qwen3-32B) under the assumption that the latter would provide diverse, orthogonal insights. However, in the high-stakes context of rare disease diagnosis, the reasoning gap between a 70B-parameter model and a 32B-parameter model proved too vast to bridge. Rather than correcting errors, the weaker models introduced "diagnostic noise"—hallucinations or common-but-incorrect heuristic biases—which, even when weighted lower, diluted the "signal" provided by the stronger model. A weighted average of a correct diagnosis (Top 1) and a hallucinated diagnosis (Missed) inevitably results in a lower confidence score than the correct diagnosis alone. This suggests that for

clinical reasoning, a coherent, high-fidelity "chain of thought" from a single, highly capable model is superior to the fragmented consensus of a heterogeneous ensemble.

Implications for Health Equity and "Lean AI"

While the failure of the multi-agent system to outperform the single model is a negative technical result, it is a profoundly positive finding for global health equity and implementation science.

The primary barrier to deploying AI in grassroots healthcare—such as rural clinics in India, Sub-Saharan Africa, or rural China—is resource constraints [36–38]. The challenges of deploying AI in LMICs are well documented: fragmented digital infrastructure, clinician shortages, data scarcity, and severe budgetary constraints [22]. Open-source, locally deployable models such as DeepSeek have been proposed as viable alternatives to cloud-dependent proprietary systems, offering transparency, economic accessibility, and local customization [22].

Multi-agent systems utilise significantly more computational power, increase latency, and introduce fragility into the software stack [39,40]. If our study had confirmed the hypothesis that only a massive, complex cluster of interacting agents could accurately diagnose complex diseases, the technology would remain inaccessible to the majority of the world's primary care facilities.

Instead, our data confirms that a Single Strong Model (e.g., GPT-oss-20b or Llama 3.1-70B), deployed locally on standard hardware, provides the optimal balance of performance and efficiency. It offers diagnostic recall and safety profiles statistically indistinguishable from, or superior to, complex ensembles. This validates a "Lean AI" strategy: healthcare systems can prioritise investing in one robust, localised model rather than orchestrating complex swarms [41]. This reduction in technical debt and hardware costs lowers the barrier to entry, making high-quality diagnostic support feasible for offline, resource-constrained environments. This aligns with emerging efforts to deploy lightweight, offline-capable LLMs in low-resource environments, such as edge-based diagnostic assistants that maintain robust reasoning without cloud dependency [12].

The "Hallucination Surface Area"

Safety is the most critical barrier to AI adoption in medicine. Our findings introduce the concept of the "Hallucination Surface Area." Every additional agent introduced into a workflow represents a new potential source of error [42–44]. This concern is amplified by reports that medical AI systems exhibit hallucination rates of 8–23%, and that multi-agent frameworks designed specifically to mitigate hallucinations through counterfactual debate or adversarial verification may themselves introduce new failure modes through complex inter-agent dynamics [25, 26]. The "hallucination surface area" concept we introduce highlights that each additional agent, each routing decision, and each weighting formula represents a point of potential failure that must be validated—a validation burden that scales multiplicatively with architectural complexity.

In the Adaptive System, the complex routing logic and weighting formulas introduced new points of failure. For instance, if the gating classifier miscategorised a neurological case as "Surgery," it would route the case to a suboptimal model workflow, inducing hallucinations that the single generalist model would likely avoid. In clinical safety, simplicity facilitates validation; validating the output of one model is significantly more practical for a human clinician than deciphering the weighted voting logic of a "black box" multi-agent system.

Limitations

This study has several limitations. First, the evaluation utilised retrospective case reports. While these represent complex real-world scenarios, they may lack the noise, missing data, and inconsistencies typical of raw Electronic Health Records (EHRs). Second, the input was limited to text, excluding multimodal data such as medical imaging (CT/MRI) or raw waveform data (ECG), which are critical for comprehensive diagnosis. Third, while we validated our automated scoring against human experts, the primary evaluation relied on an LLM-based evaluator due to the dataset size, which may introduce subtle biases. Finally, our cost analysis focused on inference time and memory usage; future studies should include direct energy consumption metrics (kWh) to further quantify the environmental sustainability of "Lean AI" versus "Agentic" approaches. Furthermore, while our findings contrast with some studies reporting benefits of multi-agent systems, direct comparison is complicated by differences in task complexity, agent homogeneity, and evaluation metrics; recent work on mixed-vendor ensembles and diagnostic diversity frameworks highlights the contextual nature of ensemble effectiveness and suggests that the benefit of MAS may be highly dependent on base model selection and task specificity [23, 24].

Future Directions

Future research should pivot from increasing architectural complexity to optimising single-model performance for medical contexts. This includes fine-tuning open-source models on high-quality medical texts and developing "Small Language Models" (SLMs) capable of running on edge devices (e.g., laptops or tablets) without sacrificing reasoning capability. Recent work has demonstrated that single, well-tuned models can achieve expert-level diagnostic accuracy, suggesting that further optimization of model size, training data, and post-training alignment may close the gap with larger cloud-based systems [20, 21]. Furthermore, prospective clinical trials are needed to evaluate the impact of these "Lean AI" tools on real-world outcomes, such as time-to-diagnosis and referral accuracy in primary care settings.

Conclusion

For AI-assisted diagnosis in primary care, simpler is often better. A single, state-of-the-art, locally hosted model provides a safer, more accurate, and significantly more cost-effective solution than complex multi-agent ensembles. Adopting this "Occam’s Razor" approach offers a scientifically validated, equitable path for deploying AI diagnostic support in resource-constrained settings, ensuring that the benefits of artificial intelligence are accessible to all, regardless of digital infrastructure.

Abbreviations

AI
Artificial Intelligence
ANOVA
Analysis of Variance
CI
Confidence Interval
CoT
Chain-of-Thought
EHR
Electronic Health Record
GP
General Practitioner
LLM
Large Language Model
LMIC
Low- and Middle-Income Country
MAS
Multi-Agent System
OPEX
Operational Expenditure
SD
Standard Deviation
VRAM
Video Random Access Memory.

Declarations

Ethics approval and consent to participate

This study involved the benchmarking of AI algorithms on de-identified, retrospective case data publicly available or synthesized from medical literature. No human subjects were prospectively enrolled, and no clinical interventions were performed on patients. As such, the study was exempt from institutional review board (IRB) approval for human subject research, in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form (including individual details, images, or videos).

Competing interests

Yansheng Li is an employee of iMEDWAY Technology Co., Ltd. All other authors declare no competing interests. The authors declare that the funding organization had no influence on the study design, data collection, analysis, or publication decisions.

Funding

This work was supported by Shandong Provincial Natural Science Foundation (No.ZR2025LSW007). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of this manuscript.

Author Contribution

Conceptualization: TC, XL; Methodology: TC, NZ; Software: NZ, YL; Validation: YL, XL; Formal Analysis: NZ, TC; Investigation: TC, NZ; Resources: YL, XL; Data Curation: NZ, TC; Writing – Original Draft: TC; Writing – Review & Editing: XL, NZ; Visualization: YL, NZ; Supervision: XL; Project Administration: XL; Funding Acquisition: XL. All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgement

The authors thank the Information Network Office of Weifang People's Hospital for providing the computational infrastructure that enabled the local deployment and benchmarking of all models in this study. We also acknowledge the clinical experts who participated in the annotation and validation of the Reference Standard dataset. Generative AI statement: During the preparation of this manuscript, the authors used DeepSeek V3.1 for language polishing and grammatical refinement of the written text. The model was not used for data analysis, interpretation of results, or generation of scientific conclusions. All intellectual content, including study design, methodology, statistical analysis, and final interpretation, remained under the sole responsibility of the authors, who reviewed and verified all content prior to submission.

Data Availability

The DiagnosisArena dataset used in this study is publicly available in a GitHub repository: https://github.com/SPIRAL-MED/DiagnosisArena and is formally described with the identifier arXiv:2505.14107. The code for model evaluation and statistical analysis is available from the corresponding author upon reasonable request. No primary clinical data were collected from human subjects in this study.

References

1. Linscott KB, Cassady JA, Robin NH. Occam's razor dulled: the occurrence of multiple genetic diagnoses. Curr Opin Pediatr. 2021;33(6):545–8. PMID:34561344. https://pubmed.ncbi.nlm.nih.gov/34561344/.
2. Falkman G, Gustafsson M, Jontell M, Torgersson O. SOMWeb: a semantic web-based system for supporting collaboration of distributed medical communities of practice. J Med Internet Res. 2008;10(3):e25. PMID:18678513. https://pubmed.ncbi.nlm.nih.gov/18678513/.
3. Singh G, Sander JW. The global burden of epilepsy report: Implications for low- and middle-income countries. Epilepsy Behav. 2020;105:106949. PMID:32143127. https://pubmed.ncbi.nlm.nih.gov/32143127/.
4. Kayola G, et al. Stroke Rehabilitation in Low- and Middle-Income Countries: Challenges and Opportunities. Am J Phys Med Rehabil. 2023;102(2S):S24–32. PMID:36634321. https://pubmed.ncbi.nlm.nih.gov/36634321/.
5. Rahman MM, et al. Maternal anemia and risk of adverse birth and health outcomes in low- and middle-income countries: systematic review and meta-analysis. Am J Clin Nutr. 2016;103(2):495–504. PMID:26739036. https://pubmed.ncbi.nlm.nih.gov/26739036/.
6. Sandmann S et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. 2025;10.1038/s41591-025-03727-2. PMID:40169761 https://pubmed.ncbi.nlm.nih.gov/40169761/
7. Shool S, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inf Decis Mak. 2025;25(1):117. PMID:40082892. https://pubmed.ncbi.nlm.nih.gov/40082892/.
8. Elkin PL, et al. Semantic Clinical Artificial Intelligence vs Native Large Language Model Performance on the USMLE. JAMA Netw Open. 2025;8(4):e256359. PMID:40204080. https://pubmed.ncbi.nlm.nih.gov/40204080/.
9. Wang L et al. A multimodal LLM-agent framework for personalized clinical decision-making in hepatocellular carcinoma. Patterns (N Y). 2025;6(12):101364. PMID:39811052 https://pubmed.ncbi.nlm.nih.gov/39811052/
10. Goswami S, Pinto EP. Employing innovative evidence-backed community processes for maternal health services by Dalit women. Int J Equity Health. 2022;21(Suppl 1):166. PMID:36457006. https://pubmed.ncbi.nlm.nih.gov/36457006/.
11. Ballhausen H, et al. Privacy-friendly evaluation of patient data with secure multiparty computation in a European pilot study. NPJ Digit Med. 2024;7(1):280. PMID:39384878. https://pubmed.ncbi.nlm.nih.gov/39384878/.
12. Liu C, et al. ChatOCT: Embedded Clinical Decision Support Systems for Optical Coherence Tomography in Offline and Resource-Limited Settings. J Med Syst. 2025;49(1):59. PMID:40056215. https://pubmed.ncbi.nlm.nih.gov/40056215/.

13. Tzanis E, et al. Agentic systems in radiology: Principles, opportunities, privacy risks, regulation, and sustainability concerns. Diagn Interv Imaging. 2026;107(1):7–16. PMID:39800655. https://pubmed.ncbi.nlm.nih.gov/39800655/.
14. Salehi S, et al. Beyond Single Systems: How Multi-Agent AI Is Reshaping Ethics in Radiology. Bioeng (Basel). 2025;12(10):1100. PMID:39859826. https://pubmed.ncbi.nlm.nih.gov/39859826/.
15. Zöller N, et al. Human-AI collectives most accurately diagnose clinical vignettes. Proc Natl Acad Sci USA. 2025;122(24):e2426153122. PMID:40029621. https://pubmed.ncbi.nlm.nih.gov/40029621/.
16. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. 2021;101(4):412–22. PMID:33454715. https://pubmed.ncbi.nlm.nih.gov/33454715/.
17. DiagnosisArena. Benchmarking Diagnostic Reasoning for Large Language Models. arXiv:2505.14107. 2025. https://arxiv.org/abs/2505.14107
18. Chen X, et al. Enhancing diagnostic capability with multi-agents conversational large language models. NPJ Digit Med. 2025;8(1):117. PMID:40082662.
19. Wang X, et al. Hierarchical agent reflection for aligning LLM reasoning with clinical diagnostic processes. Health Inf Sci Syst. 2026;14(1):23. PMID:41497892.
20. Tu T et al. Towards accurate differential diagnosis with large language models. Nature. 2025;1–7. PMID:40205049.
21. AMIE study group. Towards conversational diagnostic artificial intelligence. Nature. 2025. [Preprint/In press].
22. Can DeepSeek transform healthcare. low- and middle-income countries? Equity, governance, and deployment strategies. Innov. 2025;6(12):101009.
23. Do Mixed-Vendor Multi-Agent LLMs Improve Clinical Diagnosis? arXiv preprint. 2025.
24. MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity. 2026. [Conference proceeding/Preprint].
25. Medical AI. systems face catastrophic hallucination rates (8–23%). IEEE Xplore. 2026.
26. Dialectic-Med. Mitigating Diagnostic Hallucinations via Counterfactual Adversarial Multi-Agent Debate. ACL; 2025.
27. Independent. and collaborative performance of large language models and healthcare professionals in diagnosis and triage: a systematic review and meta-analysis. NPJ Digit Med. 2026.

Table 1

Table 1 is available in the Supplementary Files section.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- table.xls
- SupplementaryMaterials.docx