Original Paper

A Multiagent Large Language Model Framework for Emergency Treatment Recommendation in Acute Ischemic Stroke: Development and Validation Study

Bicong Yan[1,2*], MD; Ruipeng Zhang[1*], PhD; Li Chen[1], MD; Xinyu Song[1], MD; Zhongzheng Cao[1], MD; Yuehua Li[1], MD, PhD

[1]Department of Diagnostic and Interventional Radiology, Shanghai Sixth People's Hospital, Shanghai, China
[2]Faculty of Medical Imaging Technology, College of Health Science and Technology, Shanghai Jiao Tong University, Shanghai, China
*these authors contributed equally

Corresponding Author:

Yuehua Li, MD, PhD
Department of Diagnostic and Interventional Radiology
Shanghai Sixth People's Hospital
No.600 Yishan Road, Xuhui District
Shanghai 200233
China
Phone: 86 18918727305
Email: liyuehua77@sjtu.edu.cn

Abstract

Background: Acute ischemic stroke (AIS) treatment selection requires rapid, guideline-concordant integration of clinical, imaging, and laboratory data, including therapeutic windows, contraindications, stroke severity, and imaging eligibility. This process is complex, expertise-dependent, and vulnerable to safety-critical errors.

Objective: This study aimed to develop and validate a structured multiagent large language model (LLM) framework for AIS decision support using real-world cases and to assess its accuracy, safety, auditability, and impact on physician decision-making, particularly among junior physicians and nonspecialists.

Methods: We developed a multiagent LLM framework that used structured outputs and guideline-based reasoning to generate treatment recommendations (intravenous thrombolysis, endovascular thrombectomy, standard medical therapy, or non-AIS, or nonstroke) and Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification. The framework was evaluated using multicenter retrospective real-world cases from 2 hospitals collected between January 2018 and March 2025, prospective cases from February to May 2025, and literature-derived challenging cases from PubMed between January 2024 and January 2025. Performance was assessed against clinical reference standards. Safety was assessed using omission and hallucination event rates, instruction adherence, and 5-point clinical safety ratings. In a prospective physician study, physicians with different seniority and specialty backgrounds made AIS treatment and TOAST classification decisions with and without LLM support. Physician-case decision-level outcomes were analyzed using a binomial generalized linear mixed-effects model accounting for physician and case effects.

Results: The final analysis included 1055 group A cases, 721 group B cases, 144 literature-derived group C cases, and 161 prospectively collected group D cases. Across representative Baichuan, Qwen, DeepSeek, and GPT models, the multiagent framework consistently improved treatment recommendation accuracy. Model-level accuracy ranges increased from 0.546-0.737 to 0.687-0.851 in group A, from 0.587-0.698 to 0.671-0.813 in group B, and from 0.507-0.646 to 0.667-0.750 in group C. TOAST classification improved overall, with cohort-level variation. Across evaluated models, the multiagent framework increased the mean clinical safety score from 3.70 to 4.01 and reduced mean hallucination and omission rates from 33.6% to 20.6% and from 38.5% to 24.5%, respectively. In the prospective physician study, LLM support increased treatment decision accuracy from 73.1% to 88.6% (odds ratio 2.86, 95% CI, 2.27-3.60; P<.001). Accuracy gains were largest among junior and nonspecialist physicians, including junior specialists (0.667 to 0.833), junior nonspecialists (0.600 to 0.846), and senior nonspecialists (0.667 to 0.850). TOAST classification performance also improved (odds ratio 3.63, 95% CI 2.85-4.64; P<.001).

Conclusions: A structured multiagent framework improved LLM performance with average improvements of 18.9% in AIS treatment recommendation and TOAST classification, while producing more structured, auditable outputs with higher safety ratings. It was associated with higher physician decision accuracy, with larger gains among less-experienced physicians, suggesting the potential to narrow expertise-related decision accuracy gaps. Prospective multicenter studies are needed to assess effects on workflow and clinical outcomes.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400092800; https://www.chictr.org.cn/showprojEN.html?proj=248894

J Med Internet Res 2026;28:e96304; doi: 10.2196/96304

Keywords: acute ischemic stroke; large language model; multiagent system; decision support; clinical safety

Introduction

Stroke remains a leading cause of disability and death worldwide, with the greatest burden borne by low- and middle-income countries [1,2]. While acute ischemic stroke (AIS) management requires rapid and precise decision-making within narrow therapeutic windows, stroke care resources and diagnostic expertise remain profoundly uneven globally. Driven by socioeconomic inequality, geographic barriers, and workforce shortages across both resource-limited and high-income settings, these disparities lead to critical delays and misdiagnoses, worsening patient outcomes and reinforcing inequities in access to timely, specialist-level care [3-7]. These systemic challenges make equitable access to evidence-based AIS management a pressing global priority. As underscored by the World Stroke Organization Global Declaration on Stroke, which calls for equitable, evidence-based stroke systems worldwide, efforts to expand manpower and training have so far yielded only modest and uneven progress [8,9].

Large language models (LLMs) have shown potential to support clinical information synthesis and decision-making, offering a possible strategy to improve the consistency and accessibility of AIS decision support [10]. LLMs have demonstrated strong capabilities in diagnostic reasoning, differential generation, and information synthesis [11-15]. However, whether LLMs can provide reliable, guideline-concordant, and clinically safe decision support for AIS treatment selection remains insufficiently established. Most existing studies are benchmark- or theory-driven, focusing on tasks such as MedQA (United States Medical Licensing Examination) that do not adequately reflect real-world, guideline-based clinical decision-making [16,17]. This gap underscores the need for systematic case-based evaluation using real-world AIS data and physician decision-making experiments.

Our framework augments LLMs with structured clinical reasoning to support guideline-concordant therapy selection and more reliable AIS decision-making. This study had 2 objectives: to determine whether a multiagent design improves LLM performance on real-world AIS decision tasks in a multicenter, multisource evaluation, and to assess its effect on physician decision accuracy in a human-AI interaction experiment involving physicians with varying seniority and specialty backgrounds. We further examined whether LLM assistance preferentially benefits less-experienced clinicians and narrows expertise-related decision gaps, thereby supporting more consistent, safe, and accessible stroke care decision support across diverse care settings.

Methods

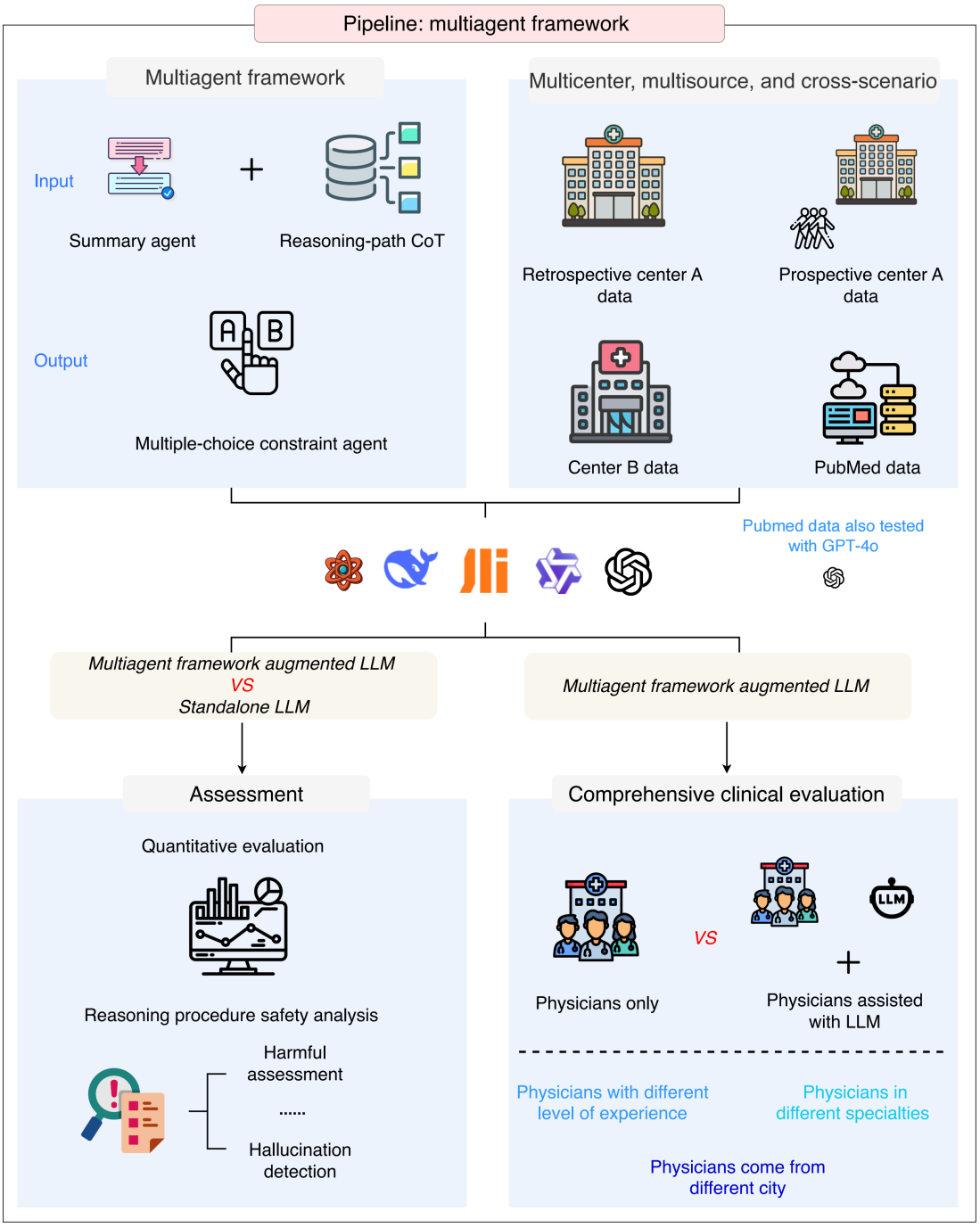
Ethical Considerations

Approval was obtained from the institutional review board of the Medical Faculty Ethics Committee of Shanghai Sixth People's Hospital affiliated to Shanghai Jiao Tong University School of Medicine (approval 2024-KY-203), and the study was registered in the Chinese Clinical Trial Registry (ChiCTR2400092800) on November 22, 2024. Informed consent was obtained from all participants. This study was conducted in accordance with the Declaration of Helsinki and relevant ethical guidelines and regulations.

Study Design

The overall study design and data sources are summarized in Figure 1. This schematic highlights the integration of multicenter retrospective and prospective real-world clinical cases with PubMed case reports, the application of the multiagent framework, and the evaluation of both model performance and human-AI interaction.

Figure 1. Study framework. The multiagent framework was benchmarked against standalone large language models (LLMs) and quantitatively validated across retrospective, prospective, multicenter, and PubMed datasets. Safety analyses encompassed harmful content, hallucinations, and numerical robustness, while clinical validation assessed human-AI interactions by comparing physicians' decisions across experience levels, specialties, and geographic locations. CoT: chain of thought.



Data Collection

We retrospectively screened clinical cases from 2 tertiary care centers: 1228 cases from center A (group A, January 2018-January 2025, tertiary grade A hospital) and 938 cases from center B (group B, May 2018-March 2025, tertiary grade B hospital). All cases were deidentified encounters of patients diagnosed with acute cerebrovascular disease. In addition, 327 stroke case reports were retrieved from PubMed between January 2024 and January 2025 (group C), which predominantly represented diagnostically challenging and clinically complex presentations. For prospective validation, 213 patients were consecutively enrolled at center A between February and May 2025 (group D). Detailed inclusion and exclusion criteria for each group are provided in the Supplementary Material Methods section and Figure S1 in Multimedia Appendix 1.

Patient Cases

To ensure patient privacy, all personally identifiable information was removed. Each case was formatted as a single paragraph containing all or a subset of the following elements: patient age and sex, chief complaint, current symptoms, medical history (including illnesses and

medications), relevant family history, physical examination findings, laboratory test results, and imaging reports.

Evaluation of Standalone LLMs

To evaluate LLM performance in generating AIS treatment recommendations and assigning stroke subtypes according to the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification, each model was tasked with generating a treatment recommendation and the corresponding TOAST classification conclusion. Seven models were tested: Baichuan-M1-14B [18], GPT-OSS-20B [19], Qwen2.5-32B [20], DeepSeek-R1-Distill-Qwen2.5-32B [21], GPT-OSS-120B [19], DeepSeek-R1-671B [21], and GPT-4o (OpenAI; only in group C dataset) [22] (Table S1 in Multimedia Appendix 1). Real-world clinical cases from groups A to D were used for this assessment. Outputs were produced in free-text format without predefined options, reflecting the probabilistic nature of clinical reasoning. To ensure consistent and reproducible evaluation, an automated grader agent was used to quantify accuracy across all LLMs (Supplementary Material Methods section in Multimedia Appendix 1).

All LLMs were evaluated in single-turn interactions using their default parameter configurations, without additional manual tuning. Model inference was performed with the official vLLM framework, providing an optimized environment for efficient large-scale deployment. All models were executed on a cluster of 8 NVIDIA H20-141GB graphics processing units (GPUs) using the official Docker release. For version control, vLLM v0.8.4 was applied to all models except GPT-OSS-20B and GPT-OSS-120B, which were deployed under v0.10.1, thereby ensuring reproducibility and transparency across experiments.

The Multiagent Framework

Overview

In addition to standalone evaluations, we implemented a multiagent framework to examine whether structured reasoning and constrained outputs could enhance model performance in AIS-specific tasks. The framework integrates three components: (1) a workflow-oriented summary agent that extracts disease-relevant evidence from lengthy clinical narratives, (2) a guideline-concordant reasoning-path chain of thought (Short-CoT) Agent that enforces structured diagnostic steps, and (3) a clinically inspired multiple-choice constraint agent that standardizes outputs within evidence-based decision boundaries.

Summarization Agent

To mitigate performance degradation caused by lengthy case tokens and to ensure the extraction of salient details, a workflow-oriented summary agent was implemented. Using DeepSeek-R1 with few-shot prompting, this agent generated structured case summaries for downstream reasoning.

Short-CoT Agent

The Short-CoT module was designed as a concise sequence of 4 critical diagnostic steps, mirroring clinical decision trees. It was developed collaboratively with neurologists, interventional radiologists, and emergency physicians by restructuring existing clinical guidelines (refer to the prompt provided in Supplementary Material Methods in Multimedia Appendix 1).

Multiple-Choice Constraint Agent

For clinically inspired final decision-making, models augmented by the multiagent framework were evaluated using 4 predefined categories as answer options: *thrombolysis, mechanical thrombectomy, standard medical therapy, and non-AIS or nonstroke conditions*. This constraint not only improved consistency but also aligned model outputs with clinically interpretable categories.

Alternative Framework Compositions

We also evaluated 3 alternative framework compositions: F1 included the Short-CoT agent and the multiple-choice constraint agent but omitted the summary agent; F2 included the guideline-derived long CoT (Long-CoT) agent and the multiple-choice constraint agent but omitted the summary agent; F3 included the summary agent and the Long-CoT agent combined with the multiple-choice constraint agent. This part used a Long-CoT agent to augment the selected LLM decision-making. The Long-CoT involved filtering and restructuring clinical guidelines into a structured decision-making pathway for AIS (refer to the prompt provided in Supplementary Material Methods in Multimedia Appendix 1) [23].

Model-Level Performance and Output Safety Assessment

Ground truth for model-level evaluation was defined as the actual treatment decision and corresponding TOAST classification documented in the clinical record, with independent verification for guideline concordance performed by 2 senior stroke clinicians: a neurologist with >10 years of experience and a neurointerventional physician with >15 years of experience. Cases with disagreement or ambiguity regarding guideline concordance were jointly re-evaluated, and eligibility for inclusion was determined by consensus.

The model-level primary outcome was the overall therapeutic and TOAST diagnostic accuracy of LLMs. Accuracy was the case-level proportion correct; TOAST F1 was macroaveraged one-vs-rest F1 across classes, and diagnostic F1 was computed on binary correctness (1/0). Model-level secondary outcomes encompassed safety-related assessments, including instruction adherence, structured harmfulness assessment, and other exploratory safety metrics. For each case, the presence of any omission or hallucination was counted as one event (yes or no), and event rates were computed as events per case (events/total cases) [24,25].

Comprehensive Multicenter, Multisource, and Cross-Scenario Evaluation

Validation Using PubMed Case Reports and External Cohorts

To clinically validate the framework, we analyzed real-world cases from group B using 6 open-source LLMs. To further test generalizability across model families, we also evaluated 6 open-source LLMs and an additional closed-source LLM in group C (PubMed case reports). The PubMed case report dataset is publicly accessible and predominantly comprises diagnostically challenging cases. This dataset therefore provided a more stringent test of the generalizability of the multiagent framework across patients with varying levels of difficulty and complexity.

Prospective Clinical Validation With Multilevel and Multispecialty Physicians

To evaluate the clinical impact of LLM assistance, we conducted a prospective, blinded physician study at center A (approval 2024-KY-203) between February and May 2025. Twelve physicians with heterogeneous AIS expertise were recruited across 4 Chinese cities or provinces (Hunan, Guangdong, Jilin, and Shanghai), comprising junior (n=5), senior (n=5), and expert strata (n=2) and including stroke specialists and nonspecialists. Consecutive eligible patients were enrolled and randomized at enrollment to an AI-assisted arm (with LLM support) or a standard review arm (without LLM support). Each case was evaluated only once by a single assigned physician.

In the AI-assisted arm, the best-performing model from the retrospective evaluation (DeepSeek-R1) generated treatment recommendations, TOAST classifications, and structured reasoning. These model outputs were provided to physicians together with the routinely available clinical materials for each assigned case. Physicians additionally provided a 5-point Likert rating of the perceived effectiveness of the AI-assisted reasoning (with higher scores indicating greater perceived usefulness). In the standard review arm, physicians reviewed identical case materials without any model output or additional computer-based decision support beyond routine hospital systems.

The reference standard was the final multidisciplinary team decision. Additional design details, including sample-size estimation and allocation procedures, are reported in the Supplementary Material Methods in Multimedia Appendix 1.

Statistical Analysis

All analyses were performed in Python (version 3.10; Python Software Foundation). Statistical significance was set at a 2-sided $P<.05$. Binary and continuous variables were summarized descriptively. Accuracy differences between standalone and LLMs augmented by the multiagent framework were tested with the McNemar test, and F_1-score differences were estimated by bootstrap resampling (1000 iterations) with 95% CIs. Paired ordinal outcomes were compared using the Wilcoxon signed-rank test. The ablation study compared the full multiagent framework with predefined variants that removed or substituted key modules (summary agent, guideline-derived Long-CoT vs Short-CoT, and free-form vs constrained multiple-choice outputs) [26]. We fitted physician-case–level binomial-logit generalized linear mixed-effects models (GLMMs), with decision correctness as the binary outcome, and LLM support as the main fixed effect for treatment decision and TOAST classification [27]. Event rates—including omission and hallucination frequencies—and differences between AI-assisted and nonassisted arms were analyzed using chi-square or Fisher exact tests, as appropriate. Outcome metrics and definitions are summarized in Table S2 in Multimedia Appendix 1.

Results

Patient Characteristics

At center A, 1228 cases were screened retrospectively, and 1055 (85.91%) were included in group A. At center B, 938 were screened, and 721 (76.87%) were included in group B. In addition, 213 prospective cases were screened at center A between February and May 2025, of which 161 (75.59%) were included in group D. Of 327 PubMed case reports retrieved, 144 (44.04%) met the inclusion criteria and were included in group C. We finally analyzed 1937 clinical cases (mean age 68.0, SD 13.6 years; n=761, 39.29% women) between January 2018 and May 2025, and 144 PubMed case reports identified from January 2024 to January 2025 (Table 1).

Table 1. Characteristics of enrolled patients.

Variables	Group A (center A; n=1055)	Group B (center B; n=721)	Group C (PubMed cases; n=144)	Group D (center A; n=161)
Age (y), mean (SD)	69.86 (13.26)	66.55 (11.76)	56.46 (18.73)	72.32 (12.22)
Male, n (%)	665 (63.03)	485 (67.27)	71[a] (50)	97 (60.25)
Female, n (%)	390 (36.97)	236 (32.73)	71[a] (50)	64 (39.75)
Disease categories, n (%)				
AIS[b]	998 (94.6)	610 (84.6)	127 (88.19)	159 (98.76)
Cerebral hemorrhage	26 (2.46)	64 (8.88)	1 (0.69)	1 (0.62)
Epilepsy	8 (0.76)	18 (2.5)	1 (0.69)	0 (0)

Variables	Group A (center A; n=1055)	Group B (center B; n=721)	Group C (PubMed cases; n=144)	Group D (center A; n=161)
Arterial aneurysm	4 (0.38)	3 (0.42)	0 (0)	1 (0.62)
Transient ischemic attack	9 (0.85)	22 (3.05)	4 (2.78)	0 (0)
Other non-AIS diseases	10 (0.95)	4 (0.55)	11 (7.64)	0 (0)
AIS treatment n (%)[c]				
Thrombolysis	135 (13.53)	135 (22.13)	17 (13.39)	20 (12.58)
Endovascular thrombectomy[d]	297 (29.76)	125 (20.49)	44 (34.56)	27 (16.98)
Standard medical management	565 (56.61)	350 (57.38)	66 (51.97)	112 (70.44)
TOAST[e] diagnosis of AIS n (%)[c]				
Large artery atherosclerosis	760 (76.15)	494 (80.98)	41 (32.28)	99 (62.26)
Cardioembolism	98 (9.82)	40 (6.56)	28 (22.04)	21 (13.21)
Small vessel occlusion	95 (9.52)	37 (6.07)	6 (4.7)	37 (23.27)
Other causes	26 (2.61)	15 (2.46)	45 (35.43)	1 (0.63)
Cryptogenic	18 (1.80)	24 (3.93)	5 (3.94)	1 (0.63)
NIHSS[f], mean (SD)	8.3 (9.1)	6.3 (7.1)	6.9 (8.3)	7.1 (8.3)
Time from symptom onset (min), median (IQR)	360.0 (180-1440)	360.0 (150-1435)	660.0 (210-1440)	360.0 (210-960)

[a]Two cases had missing sex information.
[b]AIS: acute ischemic stroke.
[c]Percentages for AIS treatment and TOAST diagnosis were calculated using the number of confirmed AIS cases in each group as the denominator (Group A: n=998; Group B: n=610; Group C: n=127; Group D: n=159).
[d]Including bridging therapy (thrombolysis and endovascular thrombectomy).
[e]TOAST: Trial of ORG 10172 in Acute Stroke Treatment.
[f]NIHSS: National Institutes of Health Stroke Scale.

Baseline Performance Disparities Among Standalone LLMs

Figure 2A and 2B and Figure S2 in Multimedia Appendix 1 illustrate the simplified system architecture for the standalone LLM and the LLMs augmented by the multiagent framework. Performance varied markedly by model size. Standalone larger-scale LLMs consistently outperformed smaller ones in both treatment recommendation and TOAST classification (Table 2). Additionally, the multiagent framework consistently improved macroaveraged sensitivity and specificity across all evaluated models (Table S3 in Multimedia Appendix 1). For example, across the pooled cohort, GPT-OSS-120B achieved 0.724 accuracy for treatment recommendation, compared with 0.570 for Baichuan-M1-14B and 0.464 for GPT-OSS-20B (all adjusted $P<.0001$; Table 2). Performance trends were consistent across individual subgroups (Table S4-S7 in Multimedia Appendix 1). Similar gaps were observed for TOAST classification, with DeepSeek-R1 surpassing all smaller models. These results establish model size as a key determinant of baseline performance.

Figure 2. Simplified system architecture and performance of standalone vs framework-augmented large language models (LLMs). (A) and (B) Simplified system architecture: constrained inputs use <think>, <treatment>, and <diagnosis> tags. (C-H) Accuracy of acute ischemic stroke (AIS) treatment recommendation for 6 LLMs (Baichuan-M1-14B, GPT-OSS-20B, Qwen2.5-32B, DeepSeek-R1-Distill-Qwen-32B, GPT-OSS-120B, and DeepSeek-R1-671B) across groups A-C; paired bars compare standalone with multiagent, showing consistent gains. In group C, GPT-4o gained accuracy in treatment recommendation (+14.9%, 0.750 vs 0.653). TOAST: Trial of ORG 10172 in Acute Stroke Treatment. *P<.05, **P<.01, ***P<.001.

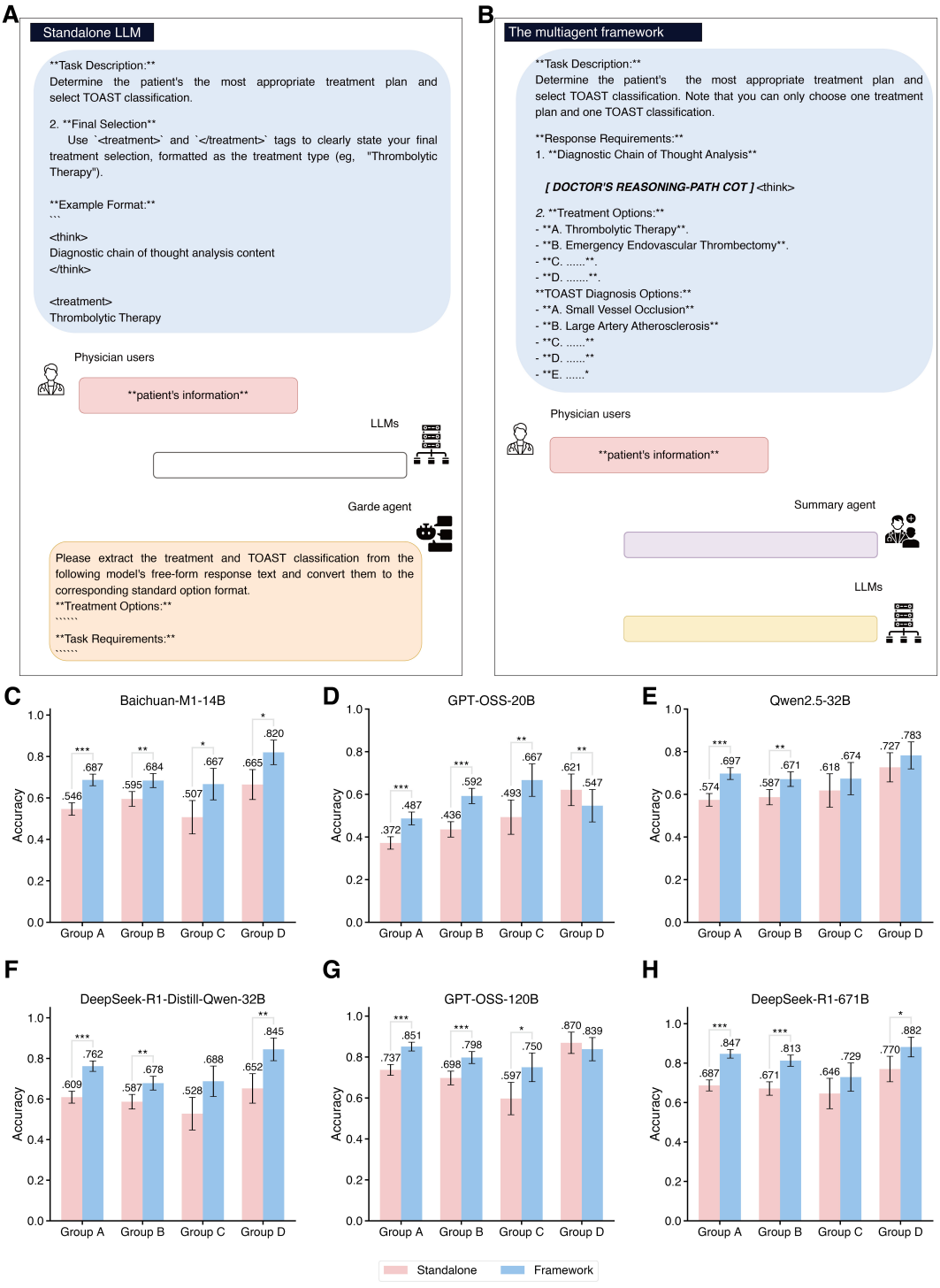


Table 2. Accuracy of large language models (LLMs) for treatment recommendation and Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification in the pooled group.

Models and accuracy[a]	Standalone LLM, accuracy	Multiagent framework, accuracy	*P* value
Baichuan-M1-14B			
Treatment	0.570 (0.547-0.591)	0.695 (0.675-0.715)	<.001

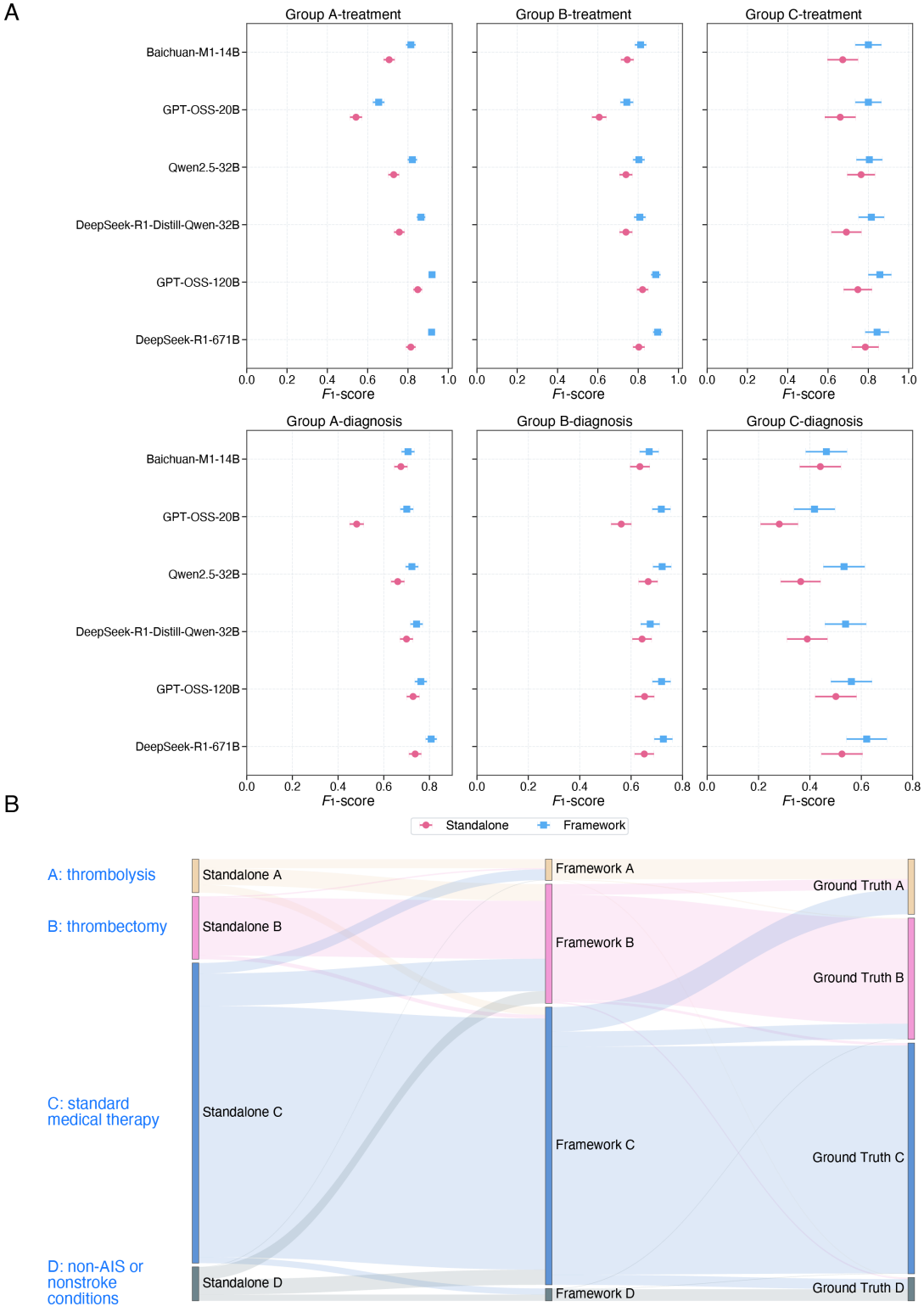
Models and accuracy[a]	Standalone LLM, accuracy	Multiagent framework, accuracy	*P* value
TOAST	0.623 (0.602-0.645)	0.657 (0.637-0.677)	.10
GPT-OSS-20B			
Treatment	0.464 (0.441-0.486)	0.541 (0.519-0.56)	<.001
TOAST	0.530 (0.507-0.553)	0.623 (0.602-0.645)	<.001
Qwen2.5-32B			
Treatment	0.593 (0.574-0.615)	0.693 (0.674-0.713)	<.001
TOAST	0.645 (0.622-0.666)	0.706 (0.686-0.726)	<.001
DeepSeek-R1-Distill-Qwen-32B			
Treatment	0.599 (0.577-0.620)	0.734 (0.716-0.754)	<.001
TOAST	0.643 (0.622-0.663)	0.689 (0.669-0.708)	<.001
GPT-OSS-120B			
Treatment	0.724 (0.706-0.744)	0.825 (0.808-0.840)	<.001
TOAST	0.690 (0.668-0.710)	0.711 (0.690-0.732)	<.001
DeepSeek-R1-671B			
Treatment	0.685 (0.667-0.704)	0.830 (0.813-0.847)	<.001
TOAST	0.680 (0.659-0.701)	0.758 (0.738-0.777)	<.001

[a]Values are presented as point estimates, with 95% confidence intervals shown in parentheses.

Multiagent Framework-Induced Improvements Across LLMs

Augmentation with the multiagent framework substantially improved outcomes across all models (Figure 2C-2H and Figure 3A; Figure S3 in Multimedia Appendix 1), with an average improvement of 18.9% compared with standalone LLMs. In group A, Baichuan-M1-14B accuracy increased by 25.8% (F_1-score +15.1%), while DeepSeek-R1-671B improved more modestly (+23.3% accuracy and F_1-score +12.7%). For TOAST classification, GPT-OSS-20B accuracy rose by 39.0% (F_1-score +4.9%), compared with only 2.5% (F_1-score −2.1%) for GPT-OSS-120B. Notably, GPT-OSS-20B exhibited unstable outputs with fluctuating gains. Collectively, these findings show that while model size remains critical for baseline accuracy, the multiagent framework reduces size-related disparities, enabling smaller models to approach the clinical utility of their larger counterparts.

Figure 3. F_1-score comparison of large language models (LLMs) in acute ischemic stroke (AIS) treatment recommendation and Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification (A) and treatment recommendation flows (B). (A) F_1-scores are shown for AIS treatment recommendation and TOAST classification across groups A-C and 6 LLMs of increasing scale. Within each LLM, paired dots represent standalone performance (orange) and framework-augmented LLMs (blue), with error bars indicating CIs. The F_1-score highlights the enhanced diagnostic reliability of the multiagent framework. (B) Sankey diagram comparing treatment recommendation flows between the standalone and the multiagent framework. Asterisks denote within-group differences between conditions, 2-sided paired tests. *P<.05, **P<.01, ***P<.001.

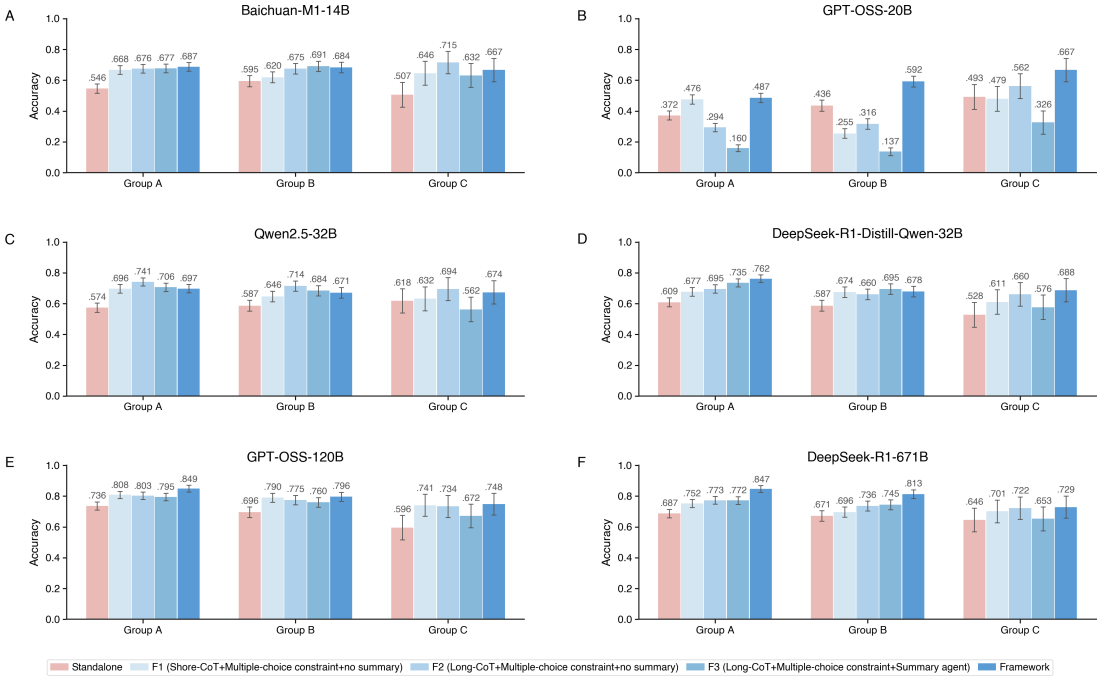


Cross-Group and Multisource Validation

Validation in groups B and C demonstrated consistent performance gains across all groups (Figure 2C-2H and Figure 3A; Tables S4-S7 in Multimedia Appendix 1). The framework-augmented DeepSeek-R1-671B achieved the highest accuracy in both treatment recommendation and TOAST classification. In group C, GPT-4o gained accuracy in treatment recommendation (+14.9%, 0.750 vs 0.653) but showed a decline in TOAST classification (−6.7%, 0.486 vs

0.521), reflecting the predominance of "other causes" and cryptogenic subtypes. The results for alternative framework compositions are shown in Figure 4. Collectively, these findings confirm that the framework provides reliable gains across heterogeneous cohorts.

Figure 4. Performance of standalone large language models (LLMs), 3 alternative framework compositions, and our multiagent framework**.** Panels A-F show accuracy for acute ischemic stroke (AIS) treatment recommendation across groups A-C. Paired bars summarize accuracy profiles across settings, providing a direct comparison of different framework compositions.



Contribution of Framework Components

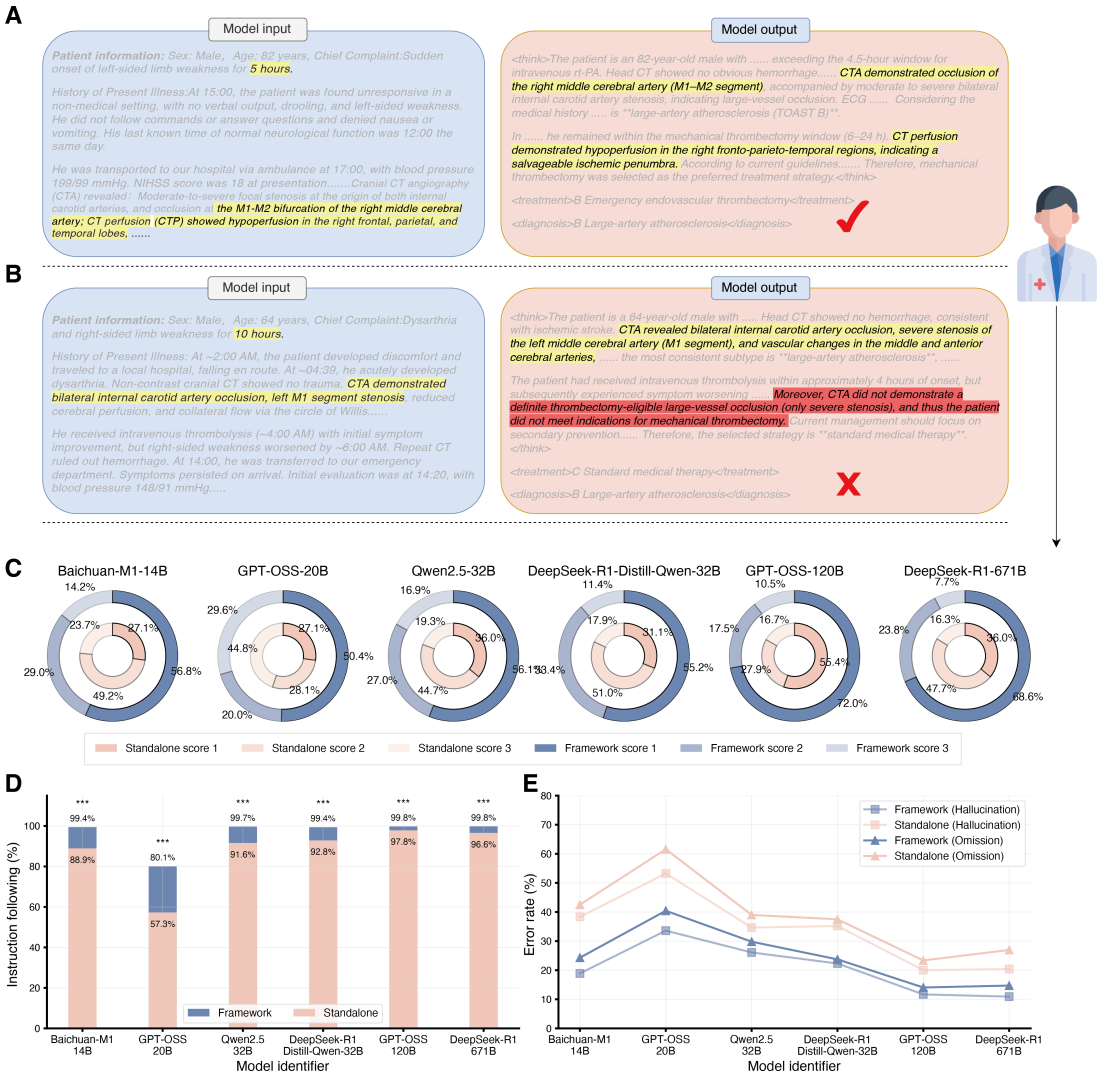
Generalized linear mixed model analysis in group A confirmed that the multiagent framework was independently associated with higher accuracy in AIS treatment recommendations (odds ratio [OR] 1.72, 95% CI 1.63-1.81; $P<.001$) and increased selection of appropriate reperfusion strategies (Figure 3B).

In the ablation study (average-only aggregation), treatment recommendation performance improved when using a multiple-choice constraint compared with free-form answer formats (constrained vs free-form: +0.038), and CoT prompting outperformed direct prompting (long vs direct: +0.052; short vs direct: +0.063), suggesting that treatment recommendations benefit from structured answer formats and explicit reasoning scaffolds. Summarization yielded only a modest average gain for treatment recommendation (summary vs no summary: +0.006). In contrast, diagnosis benefited more from summarization (+0.041) and performed better with free-form answer formats (constrained vs free-form: –0.051), motivating task-specific choices of prompting strategy and answer-format control.

Qualifying LLM Reasoning for Safe Clinical Deployment

Beyond accuracy, we visualized step-by-step reasoning traces for representative success and failure cases to characterize failure modes and deployment-relevant clinical safety risks (Figure 5A and 5B). Because accuracy alone provides an incomplete view of clinical applicability, we further assessed output safety. Compared with the standalone LLM, the multiagent framework achieved higher mean clinical safety scores (4.01, SD 0.35 vs 3.70, SD 0.43) and lower mean hallucination (20.6% vs 33.6%) and omission rates (24.5% vs 38.5%; Figure 5C-5E; Table S8 in Multimedia Appendix 1). Other exploratory safety metrics are shown in Table S9 in Multimedia Appendix 1. These findings demonstrate that structured reasoning may improve reliability and mitigate the risk of unsafe content entering clinical workflows.

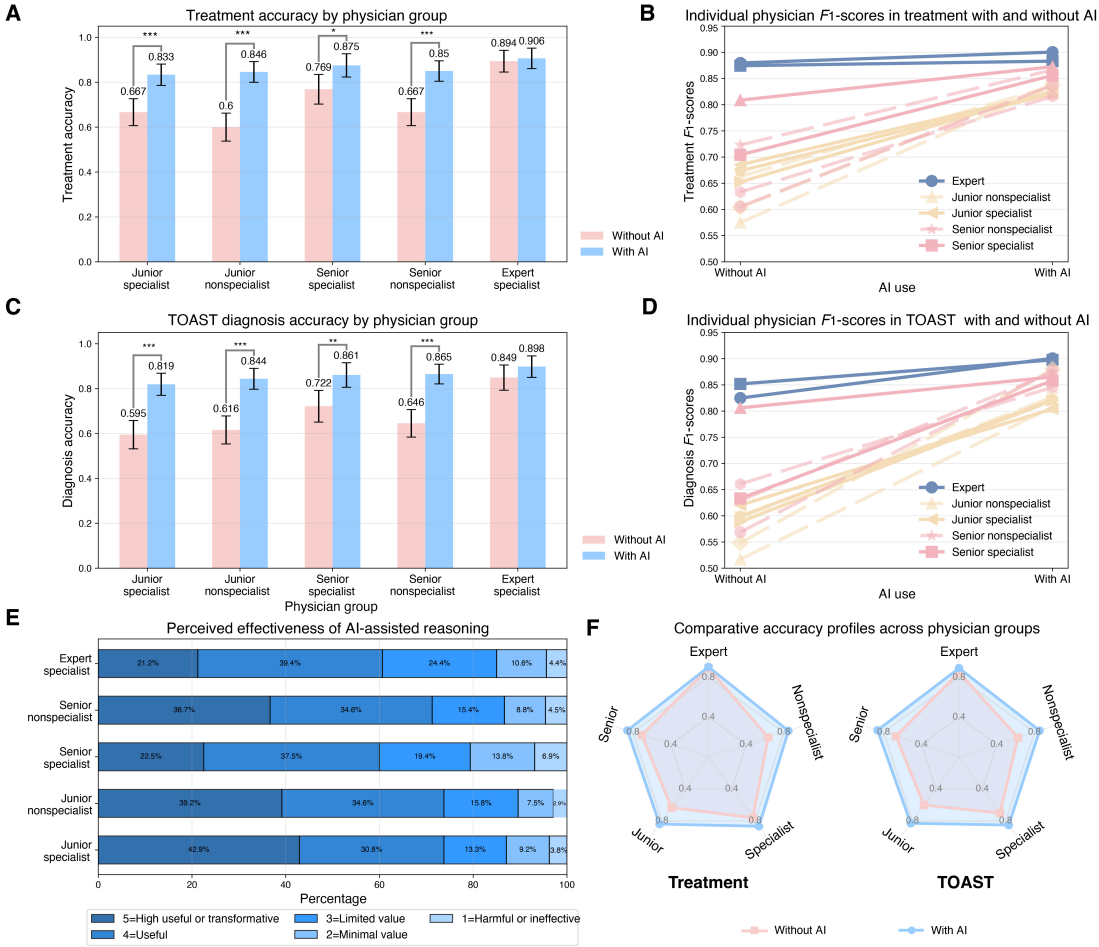
Figure 5. Case examples and safety profile of large language model (LLM) outputs. (A-B), Representative cases. (A) Correct recommendation with faithful reasoning and no hallucination or omission. (B) Incorrect recommendation: relevant details were identified, but hallucinated content led to an erroneous conclusion. (C-E) Safety metrics across 6 LLMs. (C) Harmfulness ratings on a 3-point scale (1=not harmful; 3=highly harmful) shown as a concentric doughnut plot with overlaid points. (D) Instruction-following compliance is shown as a bar plot of adherence proportions. (E) Incidence rates of hallucination and omission are shown as line plots. *P<.05, **P<.01, ***P<.001.



LLM Support Benefits Less-Experienced Physicians

Integration of LLM support substantially enhanced physician performance, with the most pronounced gains observed for less-experienced physicians in treatment decisions (0.667 to 0.833 in junior specialists; 0.600 to 0.846 in nonspecialists). On the basis of physician-case decision-level observations (1 observation per physician per case), we fitted a binomial-logit GLMM with decision correctness as the binary outcome, LLM support, physician experience, and specialty as fixed effects, and a random intercept for case and physician. After adjustment, LLM support was associated with higher odds of correct treatment decisions (OR 2.86, 95% CI 2.27-3.60; *P*<.0001) and correct TOAST classification (OR 3.63, 95% CI 2.85-4.64; *P*<.0001; Figure S4 in Multimedia Appendix 1). The estimated association was strongest among junior and nonspecialist physicians. For treatment decisions, the ORs were 3.96 (95% CI 2.64-5.96) for junior nonspecialists and 3.22 (95% CI 2.13-4.87) for senior nonspecialists; for TOAST classification, the corresponding ORs were 4.66 (95% CI 3.01-7.22) and 4.27 (95% CI 2.74-6.66). Expert specialists showed positive but nonsignificant estimates, consistent with a ceiling effect at higher baseline performance levels (*P*<.001). Physicians further rated the assistance positively, with a mean score of 3.849 out of 5. Improvements were more modest in senior and specialist physicians (Figure 6).

Figure 6. Human-AI evaluation across physician groups**.** (A) Treatment recommendation accuracy without AI (pink) and with AI (blue); bars show means with 95% CIs. (B) Individual physicians' treatment F_1-scores; lines connect the same physician from "without AI" to "with AI." (C) Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification accuracy by physician group, as in panel A. (D) Individual physicians' TOAST F_1-scores, as in panel B. (E) Perceived effectiveness of AI-assisted reasoning on a 5-point Likert scale (5=highly useful or transformative; 1=harmful or ineffective), stratified by physician group. (F) Radar plots summarizing accuracy profiles for treatment recommendation (left) and TOAST classification (right), with and without AI assistance. Physician groups are stratified by seniority (junior vs senior, including both specialists and nonspecialists) and by specialty background (specialist vs nonspecialist, including both junior and senior physicians). *$P<.05$, **$P<.01$, ***$P<.001$.



Discussion

This study advances the clinical translation of LLMs for AIS decision support by moving beyond benchmark-style evaluation to a multicenter, multisource assessment across retrospective, prospective, and literature-derived cases. First, the structured multiagent framework improved guideline-concordant therapy selection while producing more reliable and clinically accountable outputs, with lower hallucination and omission rates. Second, in a human-AI interaction experiment involving physicians of varying seniority and specialty backgrounds, the framework improved physician decision accuracy, with the largest gains observed among junior and nonspecialist physicians. Collectively, these findings suggest that LLM-integrated systems can enhance both intrinsic decision reliability and downstream clinical performance, offering a practical means to improve consistency, safety, and equity in AIS care when access to stroke expertise is uneven.

Global inequities in stroke care remain profound, with low- and middle-income countries constrained by limited resources and specialist expertise, while high-income countries face overcrowded emergency services and workforce pressures. Existing evaluations of LLMs have been largely confined to simplified benchmark tasks that fail to capture the real-world complexity of disease management [16,20,21]. To address this gap, we applied LLMs to real-world clinical and literature-derived cases, thereby simulating authentic clinical scenarios and reflecting heterogeneous contexts. Performance declined in complex settings, partly due to limitations in handling long token inputs [28,29]. To address this challenge, our multiagent framework incorporated a custom-designed summarization agent that extracted salient features from clinical narratives, improving accuracy across diverse scenarios. Augmented LLMs achieved accuracies ranging from 0.541 to 0.830 across model sizes—a level comparable to question-and-answer–style or simulated patient scenarios [13,30], and consistently higher than standalone LLMs.

This study represents an initial step toward translating LLM-assisted decision support for AIS into clinical practice, with a focus on reducing variability in care that arises from uneven clinical expertise. Prior work suggests that CoT prompting and fixed-answer formats can partially

address these limitations [31,32]. Our multiagent framework introduces a workflow-oriented, guideline-concordant structure that improves decision accuracy while maintaining clinically consistent recommendations, supporting feasibility for future clinical integration. Unlike approaches that rely on ever-larger proprietary models [33], our results show that workflow-inspired structured design improvements can yield substantial gains without increasing model scale, potentially lowering adoption barriers and supporting broader access to specialist-level decision support.

Our evaluation offers one of the most comprehensive simulations of clinical practice to date, establishing a foundation for the deployment of LLMs in AIS workflows. While the promise is substantial, deployment at scale carries risks of unintended harmful consequences [34,35]. By integrating data from multicenter, multitier hospitals, retrospective and prospective real-world cases, literature-derived high-difficulty cases, and human-AI interactions across physicians of different levels and specialties, our evaluation captured the heterogeneity and complexity of AIS care. Notably, LLMs provided the most benefit to less-experienced physicians, narrowing expertise gaps across experience, specialty, and geography, and aligning with prior reports of near expert-level performance [36,37]. In alignment with the World Stroke Organization's Global Stroke Declaration, which underscores that quality stroke care should be universal [2,38], our results advance the case for AI-enabled strategies to promote equity in global stroke systems.

This study systematically assessed safety, a critical prerequisite for clinical deployment. Because LLMs predict the next token without verifying evidence, they remain prone to hallucinations that can erode trust and generate harmful or misleading recommendations [39,40]. In our evaluation, hallucinations and omissions were not eliminated but occurred at relatively low frequencies, with the multiagent DeepSeek-R1-671B achieving a hallucination rate of 10.9%, an omission rate of 14.7%, and an overall clinical safety score of 4.36 out of 5. These findings indicate that while safety concerns remain, the error rates are within a range that may be acceptable for decision support use, supporting the feasibility of cautious clinical integration [41]. The multiagent-augmented LLMs further demonstrated stronger instruction adherence, lower hallucination and omission rates, and more reliable structured outputs, thereby addressing a key barrier to real-world deployment.

Our results suggest that structured multiagent LLM frameworks may improve guideline-concordant AIS decision support, output safety, and auditability, but the remaining safety-critical errors highlight the need for caution before clinical deployment. Such systems should be used only as human-in-the-loop decision-support tools, with final treatment decisions remaining under physician responsibility. Future workflow-based studies should prospectively test safeguards, including uncertainty signaling, guideline and local-protocol alignment, contraindication review, deferral under missing or ambiguous information, high-risk warnings, and complete logging of model inputs, outputs, and reasoning traces.

This study has several limitations. First, the framework was evaluated in case-based rather than real-time emergency room settings, precluding assessment of usability, clinician trust, latency, time to decision, bedside integration, patient outcomes, and uncertainty escalation. Second, the exclusion criteria may have produced a more standardized dataset than real-world emergency stroke care. By excluding cases with incomplete information, chronic-phase presentations, competing urgent conditions, or treatment refusal or discontinuation, the study may underrepresent information-limited and operationally constrained scenarios. Third, although the framework reduced omission and hallucination events, the proposed deployment guardrails were not prospectively tested in real clinical workflows, limiting conclusions about operational safety. Fourth, the automated grader used to categorize free-text standalone LLM outputs may have introduced evaluation error, particularly for ambiguous or internally inconsistent responses. Finally, rapid iteration of commercial LLMs and restricted access to proprietary systems may affect reproducibility and long-term stability. Future prospective multicenter workflow studies should evaluate the framework under more complex real-world conditions, with predefined safety guardrails, human adjudication of uncertain outputs, and patient-level outcome assessment.

In conclusion, our multicenter evaluation demonstrates that a structured, multiagent LLM framework significantly enhances guideline-concordant AIS decision-making (average accuracy gain: 18.9%) and output safety. Crucially, LLM support substantially increased the odds of correct physician decisions for both treatment (OR 2.86, 95% CI 2.27-3.60) and TOAST classification (OR 3.63, 95% CI 2.85-4.64), with the most pronounced benefits among less-experienced and nonspecialist clinicians. These findings highlight the framework's potential to narrow expertise gaps and support equitable stroke care, warranting further prospective evaluations of real-world workflows and patient outcomes before clinical deployment.

Acknowledgments

This study was conducted over an extended period and required substantial human and material resources. The authors thank all patients and their families for their participation, as well as the open-source community for making large language models (LLMs) publicly available. The authors are grateful to their cooperators for assistance with data collection and coding, and to the biostatistics experts for their valuable guidance on statistical methodology. They also acknowledge the strong support in clinical validation from Tao Wang, Hongmei Song, Daqian Zhang, Yingying Lu, Tonglei Fang, Xingxing Sun, Lu Fei, Yixiao Tang, Yifan Tu, Zhongzheng Cao, Fasheng Peng, Mengfan Yan, and Yuxiang Zhou.

During the preparation of this manuscript, generative AI tools, including ChatGPT (OpenAI), were used solely for language editing and linguistic polishing to improve clarity and readability. No generative AI tools were used for study design, data collection, data analysis, interpretation of results, or generation of scientific content. All scientific content, results, and conclusions were developed independently by the authors, who take full responsibility for the integrity and accuracy of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (8225024), the Key R&D subproject of the Ministry of Science and Technology (2023YFF1204804 and No. 2023YFF1204804), the Shanghai Pudong New Area Science and Technology Commission Project (PKJ2023-Y53), the Shanghai Jiaotong University, Medicine and engineering interdisciplinary program (YG2024LC08), the Shanghai key discipline of medical imaging (2017ZZ02005 and No. 2024ZZ1011), and Drug and instrument program of Shanghai Science and Technology (24SF1903900). The funders had no involvement in the study design, data collection, analysis, interpretation of results, or writing of the manuscript.

Data Availability

The raw data supporting the findings of this study are available from the corresponding author upon reasonable request. For the PubMed case cohort, the original data are not directly shared in this work; instead, we provide the references to the corresponding open-access publications in our released code repository, from which the source data can be obtained.
All code for this study is publicly available. The source code for model deployment, inference scripts, prompts, and trained model weights used in this work are available in the released code repository [42]. The following LLMs were evaluated: Baichuan-M1-14B, GPT-OSS-20B, Qwen2.5-32B, DeepSeek-R1-Distill-Qwen-32B, GPT-OSS-120B, DeepSeek-R1-671B, and GPT-4o.

Authors' Contributions

Conceptualization: BY
Data curation: BY, XS, ZC
Formal analysis: RZ
Funding acquisition: YL
Investigation: BY, LC, XS
Methodology: BY, RZ, LC
Project administration: BY
Software: RZ
Supervision: YL
Validation: ZC, LC
Writing—original draft: BY
Writing—review and editing: BY, RZ, YL

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary results, including study cohort flow and additional analyses.
[DOCX File (Microsoft Word File), 1865 KB-Multimedia Appendix 1]

References

1. GBD 2021 Forecasting Collaborators. Burden of disease scenarios for 204 countries and territories, 2022-2050: a forecasting analysis for the Global Burden of Disease Study 2021. Lancet. May 18, 2024;403(10440):2204-2256. [doi: 10.1016/S0140-6736(24)00685-8] [Medline: 38762325]
2. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization: Global Stroke Fact Sheet 2025. Int J Stroke. Feb 2025;20(2):132-144. [doi: 10.1177/17474930241308142] [Medline: 39635884]
3. Shen YC, Sarkar N, Hsia RY. Structural inequities for historically underserved communities in the adoption of stroke certification in the United States. JAMA Neurol. Aug 1, 2022;79(8):777-786. [doi: 10.1001/jamaneurol.2022.1621] [Medline: 35759253]
4. Pandian JD, Kalkonde Y, Sebastian IA, Felix C, Urimubenshi G, Bosch J. Stroke systems of care in low-income and middle-income countries: challenges and opportunities. Lancet. Oct 31, 2020;396(10260):1443-1451. [doi: 10.1016/S0140-6736(20)31374-X] [Medline: 33129395]
5. Avasarala J, Wesley K. Optimization of acute stroke care in the emergency department: a call for better utilization of healthcare resources amid growing shortage of neurologists in the United States. CNS Spectr. Aug 2018;23(4):248-250. [doi: 10.1017/S109285291700013X] [Medline: 28209212]

6. Wang H, Yu X, Guo J, et al. Burden of cardiovascular disease among the Western Pacific region and its association with human resources for health, 1990-2021: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Reg Health West Pac. 2024;51:101195. [doi: 10.1016/j.lanwpc.2024.101195] [Medline: 39286450]
7. Nasreldein A, Walter S, Mohamed KO, et al. Pre- and in-hospital delays in the use of thrombolytic therapy for patients with acute ischemic stroke in rural and urban Egypt. Front Neurol. 2022;13:1070523. [doi: 10.3389/fneur.2022.1070523] [Medline: 36742046]
8. Feigin VL, Roth GA, Naghavi M, et al. Global burden of stroke and risk factors in 188 countries, during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet Neurol. Aug 2016;15(9):913-924. [doi: 10.1016/S1474-4422(16)30073-4] [Medline: 27291521]
9. Global declaration on stroke. World Stroke Organization. 2023. URL: https://www.world-stroke.org/news-and-blog/news/global-declaration-on-stroke-commitments-for-facing-stroke [Accessed 2025-09-18]
10. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. Mar 2025;31(3):932-942. [doi: 10.1038/s41591-024-03416-6] [Medline: 39779927]
11. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. Jul 3, 2023;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
12. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. May 1, 2024;184(5):581-583. [doi: 10.1001/jamainternmed.2024.0295] [Medline: 38557971]
13. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature. Jun 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
14. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature. Jun 2025;642(8067):451-457. [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
15. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. Oct 1, 2024;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
16. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med. Apr 2025;31(4):1233-1238. URL: https://pubmed.ncbi.nlm.nih.gov/39910272/? [doi: 10.1038/s41591-024-03456-y] [Medline: 39910272]
17. Kottlors J, Hahnfeldt R, Görtz L, et al. Large language models-supported thrombectomy decision-making in acute ischemic stroke based on radiology reports: feasibility qualitative study. J Med Internet Res. Feb 13, 2025;27:e48328. [doi: 10.2196/48328] [Medline: 39946168]
18. Wang B, Zhao H, Zhou H, Song L, Xu M, Cheng W, et al. Baichuan-M1: pushing the medical capability of large language models. arXiv. Preprint posted online on Feb 18, 2025. [doi: 10.48550/arXiv.2502.12671]
19. Agarwal S, Ahmad L, Ai J, Altman S, Applebaum A, Arbus E, et al. gpt-oss-120b & gpt-oss-20b model card. arXiv. Preprint posted online on Aug 8, 2025. [doi: 10.48550/arXiv.2508.10925]
20. Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, et al. Qwen2.5 technical report. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.15115]
21. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
22. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
23. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. Dec 2019;50(12):e344-e418. [doi: 10.1161/STR.0000000000000211] [Medline: 31662037]
24. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). Aug 2, 2025;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]
25. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. May 13, 2025;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
26. Li R, Wang X, Berlowitz D, Mez J, Lin H, Yu H. CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. NPJ Digit Med. Aug 24, 2025;8(1):541. [doi: 10.1038/s41746-025-01940-4] [Medline: 40849361]
27. Hedeker D. A mixed-effects multinomial logistic regression model. Stat Med. May 15, 2003;22(9):1433-1446. [doi: 10.1002/sim.1522] [Medline: 12704607]

28. Li T, Zhang G, Do QD, Yue X, Chen W. Long-context LLMs struggle with long in-context learning. arXiv. Preprint posted online on Apr 2, 2024. [doi: 10.48550/arXiv.2404.02060]
29. Zhang G, Xu Z, Jin Q, et al. Leveraging long context in retrieval augmented language models for medical question answering. NPJ Digit Med. May 2, 2025;8(1):239. [doi: 10.1038/s41746-025-01651-w] [Medline: 40316710]
30. Wang C, Wang F, Li S, et al. Patient triage and guidance in emergency departments using large language models: multimetric study. J Med Internet Res. May 15, 2025;27:e71613. [doi: 10.2196/71613] [Medline: 40374171]
31. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. Jan 2025;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
32. Zhong W, Liu Y, Liu Y, et al. Performance of ChatGPT-4o and four open-source large language models in generating diagnoses based on China's rare disease catalog: comparative study. J Med Internet Res. Jun 18, 2025;27:e69929. [doi: 10.2196/69929] [Medline: 40532199]
33. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. Oct 2024;30(10):2886-2896. [doi: 10.1038/s41591-024-03139-8] [Medline: 39030266]
34. Habib AR, Lin AL, Grant RW. The Epic Sepsis Model falls short-the importance of external validation. JAMA Intern Med. Aug 1, 2021;181(8):1040-1041. [doi: 10.1001/jamainternmed.2021.3333] [Medline: 34152360]
35. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. Aug 1, 2021;181(8):1065-1070. [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
36. Owens D, Nguyen DQ, Dohopolski M, Rousseau JF, Peterson ED, Navar AM. Accuracy of large language models to identify stroke subtypes within unstructured electronic health record data. Stroke. Oct 2025;56(10):2966-2975. [doi: 10.1161/STROKEAHA.125.051993] [Medline: 40709446]
37. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. Apr 2024;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
38. Lee JT, Li VC, Wu JJ, et al. Evaluation of performance of generative large language models for stroke care. NPJ Digit Med. Jul 29, 2025;8(1):481. [doi: 10.1038/s41746-025-01830-9] [Medline: 40730644]
39. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care. Apr 18, 2023;27(1):148. [doi: 10.1186/s13054-023-04425-6] [Medline: 37072798]
40. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. Jun 2024;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
41. Williams CY, Bains J, Tang T, et al. Evaluating large language models for drafting emergency department encounter summaries. PLOS Digit Health. Jun 2025;4(6):e0000899. [doi: 10.1371/journal.pdig.0000899] [Medline: 40526634]
42. Enhancing clinical decision-making in acute ischemic stroke with a hybrid-reasoning framework based on large language models. Anonymous GitHub. URL: https://anonymous.4open.science/r/HR-LLM-Stroke/README.md [Accessed 2026-07-27]

Abbreviations

AIS: acute ischemic stroke
CoT: chain of thought
GLMM: generalized linear mixed-effects model
GPU: graphics processing unit
LLM: large language model
Long-CoT: guideline-derived long chain of thought
OR: odds ratio
Short-CoT: reasoning-path chain of thought
TOAST: Trial of ORG 10172 in Acute Stroke Treatment

Edited by Ivan Steenstra; peer-reviewed by Harish Vundavalli, Hruday Vairagade, Oluwatobilola Ogunbowale; submitted 27.Mar.2026; final revised version received 23.Jun.2026; accepted 23.Jun.2026; published 30.Jul.2026

Please cite as:
Yan B, Zhang R, Chen L, Song X, Cao Z, Li Y
A Multiagent Large Language Model Framework for Emergency Treatment Recommendation in Acute Ischemic Stroke: Development and Validation Study
J Med Internet Res 2026;28:e96304
URL: https://www.jmir.org/2026/1/e96304

doi: 10.2196/96304