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.

The Application Landscape Analysis of Multi-agent Collaboration Research in the Medical Field

guangneng zhu
North Sichuan Medical College
Ren Li
North Sichuan Medical College
Zining Luo
North Sichuan Medical College
Xinyu Xu
North Sichuan Medical College
Hexian Zhang
The University of Hong Kong
Jiahui Guan
The University of Hong Kong
Shipeng Hu
Peking University School and Hospital of Stomatology
Sukun Tian
Peking University School and Hospital of Stomatology
Youwen Yang
yangyouwen@jxust.edu.cn

Jiangxi University of Science and Technology
Hussein ALMasri
halmasri81@staff.alquds.edu

Al-Quds University
Junwen Wang
junwen@hku.hk

The University of Hong Kong

Method Article

Keywords:

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10668337/v1

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

Additional Declarations: The authors declare no competing interests.

The Application Landscape Analysis of Multi-agent Collaboration Research in the Medical Field

Authors:

Guangneng Zhu[1], Ren Li[1], Zining Luo[1], Xinyu Xu[1], Hexian Zhang[2], Jiahui Guan[2], Shipeng Hu[3], Sukun Tian[3], Youwen Yang[4#], Hussein ALMasri[5#], Junwen Wang[2#]

# These authors contributed equally to this article and should be recognized as the corresponding authors.

Author information

[1]North Sichuan Medical College, Sichuan, China.
[2]The University of Hong Kong, Hong Kong, China.
[3]Peking University School and Hospital of Stomatology, Beijing, China.
[4]Jiangxi University of Science and Technology, Ganzhou, China.
[5]Al-Quds University, Jerusalem, Palestine.

Corresponding Author

Youwen Yang: yangyouwen@jxust.edu.cn
Hussein ALMasri: halmasri81@staff.alquds.edu
Junwen Wang: junwen@hku.hk

Authors' contributions

GN.Z. contributed to the conception of the study, study design, review and supervision during the study, data analysis and interpretation, and drafting of the manuscript; R.L. and SK.T. contributed to the conception of the study, study design, review during the study, and data analysis and interpretation; ZN.L., XY.X., HX.Z., H.A., JH.G. and SP.H. contributed to the creation of figures and the collection of supplementary materials; JW.W., YW.Y. contributed to project management, supervision, fundraising, and review. All authors participated in the critical revision of the manuscript and approved the submission.

Ethics declarations

Competing interests

The authors declare no competing interests.

Acknowledgements

There was no funding provided for this study. We utilized ChatGPT for assistance with parts of the English proofing.

Data availability

The corresponding author had full access to all data in the study and final responsibility for the decision to submit the report for publication. The data used and analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability

The R code and Data used for the analysis is available at: https://github.com/zhuguangneng/The-Application-Landscape-Analysis-of-Multi-agent-Collaboration-Research-in-the-Medical-Field

The code for reproducing the economic benefits can be found at: https://drive.google.com/drive/folders/1bqiNF6VDhqrw2sdBXoOr9JorpLEDO2q7?usp=drive_link

List of abbreviations:

AD: Alzheimer's disease

CHA: Chatbot Health Advice

CHART: Chatbot Assessment Reporting Tool

EHR: Electronic Health Records

JCR: Journal Citation Reports

LLM: large language model

MedQA: medical question answering

OR: odds ratio

Abstract

Background: Multi-agent collaboration systems are increasingly being applied in the medical field, yet a systematic synthesis of the overall research landscape in this area remains lacking.

Objective: To systematically characterize the application landscape of multi-agent collaboration research in medicine, including publication trends, task type distribution, quantitative effects in medical question answering (MedQA) tasks, methodological reporting quality, and cost characteristics.

Methods: We systematically searched PubMed, Embase, Cochrane Library, Web of Science, and China National Knowledge Infrastructure from June 2018 to May 16, 2026. Studies applying multi-agent collaboration frameworks to medical or health-related domains were included. Bibliometric methods were used to analyze publication trends, country distribution, and journal impact. Meta-analysis was performed for MedQA tasks using odds ratio (OR) as the effect measure. We also conducted an exploratory study to replicate the comparison of economic benefits between multi-agent systems and single LLMs as described in the original paper. Reporting completeness was assessed using the Chatbot Assessment Reporting Tool (CHART) guideline. Cost data were extracted and summarized.

Results: Of 49 included articles, publications grew rapidly from Q3 2024 to Q2 2026 (23 in 2025, 23 in 2026). China (34.7%) and the United States (32.7%) contributed most articles; most appeared in JCR Q1 journals, with npj Digital Medicine publishing the most (n=5). Meta-analysis of 30 comparisons from 8 studies showed multi-agent systems showed higher performance LLMs in MedQA ($p < 0.001$). Paired subgroup analysis from the single study with item-level data confirmed the net advantage ($p < 0.001$). CHART assessment revealed systematic reporting deficiencies: full prompts (37.0%), query date/location (0%), and pre-registration (2.2%). Cost-effectiveness varied by architecture; one study favored multi-agent systems, while another favored single models.

Conclusion: Multi-agent collaboration significantly improves MedQA accuracy over single LLMs, a finding robust across sensitivity analyses. However, reporting transparency and reproducibility remain critically deficient; research is geographically concentrated in China and the United States; and cost-effectiveness is architecture-dependent. Future research should prioritize standardized reporting, prospective multicenter validation, and systematic cost-utility analyses.

Introduction

In recent years, large language models (LLMs) have shown rapid growth in medical research and applications[1]. Representative models such as ChatGPT have been extensively applied in clinical documentation, diagnostic assistance, patient communication, and medical education due to their strong language understanding and generation capabilities[2]. With the release of medical-specialized LLMs such as Med-PaLM 2, their accuracy on medical benchmarks like medical question answering (MedQA) has reached 86.5%, and some models have achieved performance approaching or exceeding human expert level on medical licensing examinations[3,4]. LLMs have been demonstrated to have application potential in clinical documentation, decision support, patient education, and workflow optimization[4].

Against this backdrop, researchers have begun to explore systems composed of multiple LLMs, known as multi-agent collaboration frameworks[5]. Park et al. introduced the concept of generative agents, designing an architecture with memory stream, reflection, and planning modules. In a simulated sandbox environment, they demonstrated spontaneous social behaviors among 25 agents, including information diffusion, relationship formation, and coordination of group activities, showing the feasibility of LLM-based agents in simulating human-like behaviors. Inspired by this work, medical researchers have started to apply multi-agent collaboration frameworks to complex tasks such as clinical reasoning[6]. These systems simulate multidisciplinary team consultations, role division, and debate mechanisms, enabling multiple agents to collaboratively complete complex medical reasoning tasks[7]. Prior studies have applied multi-agent collaboration to disease diagnosis, MedQA, imaging report generation, and electronic health record (EHR) modeling[8]. For example, a multi-agent dialogue framework was developed for rare disease diagnosis, significantly improving LLM performance on complex cases[7,9]; another framework enhanced the accuracy and reasoning depth of MedQA through multi-agent debate mechanisms[10]. However, a systematic understanding of the overall state of this emerging field is currently lacking—questions such as which specific task types multi-agent collaboration is applied to in medicine, the publication quality and academic impact of existing studies, and the actual performance across various tasks have not yet been comprehensively reviewed and quantitatively evaluated.

To this end, this study aims to conduct a systematic application landscape analysis of multi-agent collaboration research in the medical field. We will use bibliometric methods to depict the overall development profile of this field from dimensions such as publication trends and impact factors. At the same time, we will classify the included studies by task type, conduct a meta-analysis on MedQA tasks to quantitatively evaluate their performance, extract data from the articles or reproduce the

open-source articles to compare the economic benefits between multi-agent systems and single LLMs, and perform descriptive analyses for other task types. We will then evaluate the existing literature according to relevant reporting guidelines to identify limitations of current studies. Finally, we will extract cost data from articles that reported the costs of using multi-agent systems, to gain an overall understanding of the cost profile of this approach. Through these efforts, we aim to provide a comprehensive picture of the current state of multi-agent collaboration research in medicine, and to offer references for future research directions and methodological improvements.

Methods

Registration

This systematic review was prospectively registered with PROSPERO (CRD420261370231).

Definition of the multi-agent

A multi-agent system refers to a computing system composed of multiple autonomous agents, each driven by an LLM as its core engine and capable of perceiving the environment, making decisions, and executing actions. Unlike a single LLM that only generates static text responses, each agent in a multi-agent system can complete complex tasks through iterative reasoning, autonomous tool use, and inter-agent collaboration[6]. The generative agents architecture proposed by Park et al. stores complete experience records of agents in natural language, synthesizes memories into higher-level reflections over time, and dynamically retrieves them to plan behaviors, demonstrating for the first time the feasibility of LLM-driven multi-agent systems simulating believable human behaviors in open-world environment[6].
In multi-agent systems, agents enhance overall capability through structured collaboration and division of labor. For example, Virtual Lab organizes LLM agents into roles such as principal investigator, domain expert, and scientific reviewer, where agents exchange viewpoints, challenge each other, and gradually converge on research strategies through structured discussions[11]. In Agent Hospital, patients, nurses, and doctors are all LLM-driven autonomous agents that simulate the entire process from disease onset, triage, consultation, examination, diagnosis, to treatment and recovery. Tzanis et al. pointed out that the core of such systems lies in embedding LLMs into frameworks that enable reasoning, planning, and acting, allowing them to dynamically interact with users, tools, and data sources[12]. Salehi et al. further defined agentic AI operationally as requiring mandatory iterative reasoning, with autonomous tool use or multi-agent collaboration as an additional capability[13].

In summary, a multi-agent system is an autonomous computational system that uses LLMs as its reasoning engine, possesses memory and reflection capabilities, and accomplishes complex tasks through inter-agent collaboration[14]. Its core features can be summarized as three points: first, each agent uses an LLM as the core engine for reasoning and decision-making; second, agents possess memory, reflection, and planning abilities, enabling them to continuously refine their behavior based on past experience; and third, multiple agents accomplish complex tasks that a single model cannot handle through division of labor and collaboration.

Search strategy and study selection

This systematic review collected and screened studies that validated multi-agent collaboration in the medical field based on LLMs. To maximize the inclusion of eligible studies, we adopted a search strategy centered on "large language model" and its variants to ensure comprehensive coverage. Additionally, 174 mainstream model types and 35 corresponding institutional names were identified via prominent LLM leaderboards—specifically, OpenCompass (https://opencompass.org.cn/) and the Hugging Face Open LLM Leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)—and integrated into the search formula. The detailed search strategy is provided in the supplementary materials: **Retrieval Strategy**.
Since the first generative AI model (GPT-1) was released in June 2018[15], the search time frame was set from June 2018 to May 16, 2026, to cover as comprehensively as possible all studies in this field since the advent of generative pre-trained transformer models. We searched the following English electronic databases: PubMed, Embase, Cochrane Library, and Web of Science. We also searched the primary Chinese academic database: China National Knowledge Infrastructure.

All stages of the review (Apart from title and abstract screening) —including full-text review, data extraction, and risk-of-bias assessment—were performed independently and in duplicate by two reviewers. Any discrepancies were resolved through consensus or by consultation with a third independent reviewer. Data were extracted using a standardized, pre-designed template. The number of excluded studies, along with the specific reasons for exclusion, were documented in a flow diagram.

The study selection process was as follows. First, title and abstract screening was conducted. The inclusion criteria for this initial screening were: (1) the study content involved LLMs; (2) the research topic was related to medicine or health; and (3) the language was limited to English or Chinese. Studies that clearly used only a single LLM were excluded.
Screening Method: Given the need to screen over 1 million titles and abstracts, manual screening was not feasible in practice. We therefore used an LLM-assisted screening workflow[16]. Specifically, for the initial AI screening, we used [gpt-5-mini-2025-08-07](https://platform.openai.com/docs/models/gpt-5-mini) to make relevance judgments on titles and abstracts. The prompt was calibrated and then fixed. The AI output was one of three labels: "include," "exclude," or "uncertain." The prompt is provided in **Supplementary material: Section 2: Prompts**.

Reliability Verification: For the 126,822 records that the AI classified as "exclude," we randomly selected 1% (n = 1,268) for

manual re-evaluation; this process was independently repeated three times. For all 587 records classified as "include/uncertain," we randomly selected 10% (n = 59) for manual re-evaluation, also independently repeated three times. Based on the sampled data, we constructed a confusion matrix and calculated the overall agreement rate and Cohen's kappa coefficient to quantify the level of agreement between AI and human judgments. Kappa values were interpreted according to Landis et al.[17]: <0.20, slight agreement; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect agreement. (Detailed methods and results are provided in the **Supplementary Material: Section 2: Reliability Verification**.)
Discrepancy Resolution: Discrepancies were resolved through discussion or third-party arbitration.

Following the above process, the initial literature is subject to an initial screening by the LLM, after which the researchers carry out a full-text review.
For full-text inclusion, studies had to meet all of the following criteria: (1) the study adopted a multi-agent collaboration framework or approach; and (2) the research focused on specific applications in medicine or health-related fields. The exclusion criteria were: (1) studies that used only a single LLM without involving multi-agent collaboration; and (2) studies for which the full text could not be obtained. The entire screening process was recorded and managed using NoteExpress (https://www.inoteexpress.com) and Excel. The finally included studies proceeded to data extraction and quality assessment.

Extraction of basic characteristics from the included literature

To systematically describe the research landscape and distribution characteristics of this field, we extracted the following basic information from each included study: first author, publication year, article type (peer-reviewed, review, preprint, etc.), multi-agent system name, underlying LLM composition, multi-agent collaboration type, and medical task description and task category. These variables were extracted to facilitate subsequent task classification, effect stratification, and collaboration mode categorization, thereby enabling a multi-dimensional presentation of the current application landscape.

Bibliometric Methods

Data sources and indicator extraction

The following bibliometric information was extracted from each included study: publication year and quarter, journal name, impact factor (IF), CiteScore, Journal Citation Reports (JCR) quartile, and country of affiliation of the first author. All information was obtained from the article metadata and official journal websites. Impact factor and JCR quartile were retrieved based on the official data of the corresponding year for each journal.

Statistical analysis

Statistical analyses and data visualization were performed using R software (version 4.3.1). The following R packages were used: readxl, dplyr, ggplot2, lubridate, scales, and tidyr.

Publication trend analysis

Publication counts were calculated on a quarterly basis. A stacked bar chart was generated to display the number of publications contributed by each country per quarter, with a superimposed line plot showing the total quarterly publication volume. The country of affiliation is determined by the country to which the first author belongs.

Country distribution

The total number of publications per country was tallied, and the percentage contribution of each country was calculated. Countries contributing less than 3% of the total were collapsed into an "Other" category and displayed in gray.

Journal distribution

The number of articles published in each journal was recorded. A bubble plot was constructed with impact factor (log-scale) on the x-axis and CiteScore (log-scale) on the y-axis. Bubble size represented the number of articles published in that journal, and bubble color indicated the JCR quartile (Q1, Q2, Q3, Q4; journals without a quartile were classified as "Other"). The two journals with the highest publication counts were labeled directly on the plot.

Type of multi-agent collaboration

In the included studies, different multi-agent systems varied in their collaboration mechanisms. In this study, we classified multi-agent collaboration approaches into two broad categories based on the information flow and feedback mechanisms during the collaboration process: Single-Round Multi-Agent Collaboration and Multi-Round Multi-Agent Collaboration. Two reviewers independently reviewed the full texts and supplementary materials of each included study and classified the collaboration approach of each study according to the definitions below. The classification took into account the following features: whether multi-round interactions existed, whether a clear feedback mechanism was present, whether critical or revision steps were included, and whether the convergence condition was a dynamic threshold. Disagreements were resolved through discussion or by consulting a third reviewer. A comparison of the core features of the two collaboration types is provided in **Supplementary Table 1**.

Meta-analysis

Data extraction

For each included study that addressed MedQA tasks, we extracted the following data: first author, publication year, the comparator single-LLM model, and the type of multi-agent collaboration (**Supplementary Table 2**). For the meta-analysis, we extracted the number of correct responses and the total number of responses for both the multi-agent group and the comparator single-LLM group. For studies that reported accuracy as a percentage rather than raw counts, the number of

correct responses was back-calculated by multiplying the reported accuracy by the total number of cases and rounding to the nearest integer. For studies reporting data on topic-level matching, we extracted the number of cases where both were correct, where only the multi-agent system was correct, where only the single model was correct, and where both were incorrect. Studies that did not report the total number of cases were excluded.

Quality assessment

Since all included primary studies were comparative accuracy studies in which the same set of medical questions was answered separately by multi-agent systems and single LLMs to compare the accuracy between the two approaches, the ROBINS-I V2 tool (https://www.riskofbias.info/welcome/robins-i-v2) was used as the primary instrument for risk of bias assessment. This tool, recommended by the Cochrane Collaboration, is designed for assessing risk of bias in non-randomized studies of interventions and covers six domains, with each domain rated as low, moderate, serious, or critical risk of bias. The assessment results were visualized using a traffic light plot, where different colors represent the risk of bias level for each domain and for the overall assessment.

However, because ROBINS-I was designed for non-randomized intervention studies rather than AI benchmark experiments, it cannot capture biases specific to comparative evaluations of LLMs. We therefore supplemented ROBINS-I with a pre-defined, AI-specific checklist tailored to the context of MedQA benchmarks. This checklist covers eight domains derived from recent methodological recommendations for LLM evaluation[18,19]: (1) comparability of base models; (2) transparency of prompt reporting and tuning; (3) statement on test set contamination; (4) consistency of token/tool access; (5) selective reporting of best-run results; (6) reporting of randomization/replication; (7) blinding of evaluators; and (8) consistency of external knowledge/RAG use. Based on the full texts and supplementary materials of these eight studies, each domain was rated as "Yes," "No," or "Not reported."

Statistical analysis

All primary studies included in this study used a paired design, in which the same set of medical questions was answered by both a multi-agent system and a single model. However, only one study (Jiasheng Wang et al.[20]) reported question-level paired 2×2 table data, while the remaining studies reported only group marginal totals. To address this, we adopted a stratified analysis strategy. For most studies that reported only marginal totals, we used a three-level meta-analysis model to pool odds ratios (ORs) and conducted sensitivity analyses by applying variance-covariance corrections with ρ = 0.2, 0.5, and 0.8. For the only study that reported complete paired data (with four comparison pairs of model configurations), we pooled the paired 2×2 tables, calculated the pooled paired OR, and performed the exact McNemar test, as exploratory evidence for single-center internal validation.

Effect Size Calculation: In this study, we used the OR as the effect size to compare the accuracy between multi-agent systems and single LLMs in MedQA tasks. Effect sizes were calculated using the escalc function with the formula:

$$logOR = \ln\left(\frac{a/b}{c/d}\right)$$

Where a is the number of correct responses in the multi-agent group, b is the number of incorrect responses in the multi-agent group, c is the number of correct responses in the single-model group, and d is the number of incorrect responses in the single-model group. The variance was calculated as:

$$Var(logOR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Three-Level Meta-Analysis Model: Because 30 comparison pairs were nested within 8 original studies, and these comparison pairs shared study design, item source, and evaluation procedures, they did not satisfy the statistical assumption of independence. We therefore used a three-level random-effects model for data synthesis. This model partitions the total variance into three levels:

Level 1 (sampling error): sampling variation at the individual-level within each comparison pair;

Level 2 (within-study variance): heterogeneity among different comparison pairs within the same study;

Level 3 (between-study variance): heterogeneity across different studies.

The model expression is:

$$logOR_{ij} = \mu + v_i + \omega_{ij} + \varepsilon_{ij}$$

where μ is the overall pooled effect size, $v_i \sim N(0, \sigma^2_{study})$ is the random intercept for the i-th study, $\omega_{ij} \sim N(0, \sigma^2_{comparison})$ is the random intercept for the j-th comparison pair within the i-th study, and $\varepsilon_{ij} \sim N(0, \tau^2_{ij})$ is the sampling error. The random structure was specified as `random = ~ 1 | study_id / comp_id`, indicating that comparison pairs are nested within studies.

Model estimation was performed using restricted maximum likelihood. Given the small number of studies (k = 8), and to avoid inflated false positive rates that may result from conventional normal approximation under small-sample conditions, we applied the Hartung–Knapp correction for the confidence intervals and hypothesis tests of the pooled effect size. This method uses the t-distribution instead of the z-distribution, providing more conservative and robust statistical inference.

Heterogeneity Assessment: The following metrics were used to evaluate between-study heterogeneity:

Cochran's Q test: to test whether heterogeneity existed among effect sizes; a $P < 0.10$ was considered indicative of significant heterogeneity.

I^2 statistic: to quantify the proportion of total variation attributable to heterogeneity; I^2 values of 25%, 50%, and 75% were considered to represent low, moderate, and high heterogeneity, respectively.

The I^2 for the three-level model was calculated as:

$$I^2_{total} = \frac{\sigma^2_{study} + \sigma^2_{comparison}}{\sigma^2_{study} + \sigma^2_{comparison} + \tilde{v}} \times 100\%$$

where $\tilde{v}$ is the harmonic mean of the sampling variances across comparison pairs (i.e., the typical sampling variance). We also reported the between-study heterogeneity proportion (I^2_{study}) and the within-study heterogeneity proportion ($I^2_{comparison}$) separately.

Sensitivity Analysis: 1. Leave-one-out sensitivity analysis
To assess whether the pooled effect size was disproportionately influenced by any single comparison pair, we used a leave-one-out approach, sequentially removing each comparison pair one at a time, refitting the three-level model to the remaining data, and examining the range of variation in the pooled effect size. If removing any single comparison pair did not fundamentally change the direction or statistical significance of the effect size, the results were considered robust.
2. Sensitivity analysis for missing paired data (variance–covariance correction)
A key limitation of this study is that all included primary studies used a paired design, in which the same set of medical questions was answered by both a multi-agent system and a single model. However, most studies reported only marginal totals (number of correct responses / total number of responses) for each group, without providing question-level paired 2×2 table data (i.e., the counts of questions answered correctly only by the multi-agent system, correctly only by the single model, correctly by both, or incorrectly by both). Treating these marginal totals as independent samples would underestimate the standard errors and may lead to overly narrow confidence intervals.
To assess the potential impact of missing paired data on our conclusions, we kept the OR as the effect size and applied a correction to the variance-covariance matrix of the three-level model. The approach was as follows:
For any two comparison pairs i and j from the same study, if they shared the same set of medical questions (i.e., used the identical test item set), then their effect sizes $logOR_i$ and $logOR_j$ would be positively correlated. We assumed the covariance to be:

$$Cov(logOR_i, logOR_j) = \rho \cdot \sqrt{Var(logOR_i) \cdot Var(logOR_j)}$$

where ρ is the intraclass correlation coefficient. Because the true ρ could not be directly estimated from the published marginal data, we adopted a sensitivity analysis approach by setting ρ to 0.2 (weak correlation), 0.5 (moderate correlation), and 0.8 (strong correlation), with ρ = 0 as the main analysis. After constructing the full variance-covariance matrix, we refit the three-level model. If the direction and statistical significance of the pooled effect size remained consistent across the different ρ assumptions, this would indicate that the missing paired data did not introduce substantial bias to our core conclusions.

Publication Bias Assessment: The following methods were used to assess publication bias:
Funnel plot: A funnel plot was constructed with the log OR of each comparison pair on the x-axis and its standard error on the y-axis; symmetry was visually inspected.
Egger's linear regression test: This test was used to examine whether the asymmetry in the funnel plot was statistically significant; a P < 0.05 was considered indicative of publication bias.

Statistical Analysis Software: All statistical analyses were performed using R software (version 4.3.1). The following R packages were primarily used: "metafor" (version 4.6-0): for fitting three-level random-effects models (via the `rma.mv` function), leave-one-out sensitivity analysis, and publication bias tests (`regtest` and `ranktest`);
"meta" (version 7.0-1): for auxiliary effect size calculations and conventional meta-analysis;
"dplyr" and "readxl": for data manipulation and reading; "ggplot2", "ggrepel", "forestplot", and "grid": for data visualization, including forest plots, funnel plots, and leave-one-out sensitivity analysis plots.

Paired Subgroup Sensitivity Analysis (Within-Study Internal Aggregation)
Purpose and Data Source: Among the included studies, the study by Jiasheng Wang et al.[20] reported complete paired 2×2 table data under four different model configurations (claude-3.7, DeepSeek-R1, gemini-2.5-flash, and GPT-4o), comprising four comparison pairs in total.
We used a direct pooling method for paired 2×2 tables, in which the paired counts from the four comparison pairs were summed cell-wise to form a single overall paired 2×2 table, and then performed statistical inference based on this aggregated table. This approach has been widely used in single-center multi-configuration validation studies and is statistically justified: when multiple comparison pairs share the same underlying data-generating mechanism, pooling the counts can improve the precision of effect estimation[21,22] without introducing spurious bias[23,24].

Construction and pooling of paired 2×2 tables: For each comparison pair j $(j = 1,2,3,4)$, we extracted the following four counts of questions:

	Single model correct	**Single model incorrect**
Multi-agent correct	a_j (both correct)	b_j (only multi-agent correct)
Multi-agent incorrect	c_j (only single model correct)	d_j (both incorrect)

where b_j and c_j are the discordant pairs, representing the numbers of items for which the multi-agent system corrected a single-model error and for which it introduced a new error on an item that the single model had answered correctly, respectively.
Subsequently, the corresponding cells of the four comparison pairs were summed arithmetically:

$$b_{pooled} = \sum_{j=1}^{4} b_j\,, \quad c_{pooled} = \sum_{j=1}^{4} c_j$$

Effect Size Calculation and Statistical Inference: The formula for the pooled paired OR is:

$$OR_{paired} = \frac{b_{pooled}}{c_{pooled}}$$

An $OR_{paired} > 1$ indicates that the multiag-ent system has a net advantage in the paired comparison (i.e., it corrects more errors than it introduces); an $OR_{paired} < 1$ suggests a disadvantage; and an $OR_{paired} = 1$ indicates no difference in performance between the two models on discordant pairs.

Confidence interval: The variance of the log paired OR ($\ln(OR_{paired})$) was estimated using a formula based on the count of discordant pairs:

$$Var\left[\ln(OR_{paired})\right] = \frac{1}{b_{pooled}} + \frac{1}{c_{pooled}}$$

The 95% confidence interval for OR_{paired} was then calculated as:

$$exp\left[\ln(OR_{paired}) \pm 1.96 \times \sqrt{\frac{1}{b_{pooled}} + \frac{1}{c_{pooled}}}\right]$$

McNemar's exact test: To test whether the difference between the multi-agent system and the single model in discordant pairs was statistically significant, we used the exact McNemar test. Under the null hypothesis H_0 (that the two models performed equally in discordant pairs, i.e., the expected counts of b and c are equal), b_{pooled} follows a binomial distribution:

$$b_{pooled} \sim Binomial(n = b_{pooled} + c_{pooled},\ p = 0.5)$$

The two-sided p-value is calculated as follows:

$$P = 2 \times \min\left\{\Pr(X \le b_{pooled}),\ \Pr(X \ge b_{pooled})\right\}$$

where $X \sim Binomial(n, 0.5)$. When n ≥ 20, a continuity-corrected chi-square approximation may also be reported as supplementary information, but the exact test result is considered definitive.

A Comparison of the Economic Benefits of Multi-Agent Systems and Single LLMs

The economic benefit of each model in this task was calculated as the accuracy achieved divided by the financial cost incurred.

The study by Xinti Sun et al.[25] reported the costs of three single LLMs (GPT-o1, Qwen-QwQ, and DeepSeek-R1) and a multi-agent system.

In addition, only one study[26] provided accessible code and questions. In that study, for each USMLE question, five GPT-4 instances (AI_A to AI_E) were simultaneously called to independently generate responses and reasoning paths. A facilitator agent was introduced to integrate the outputs from each member. If the initial responses from the five instances (round 1, during which there was no discussion among the GPT-4 instances) showed disagreement (i.e., were not unanimous), the facilitator synthesized the reasoning basis from each side, identified key points of contention, and constructed new prompts to guide the committee into the next round of discussion. This process was iterated until all members reached consensus. Finally, the facilitator generated a synthesized conclusion as the final answer to the question. If the initial choices of the five instances were unanimous, no further rounds were conducted. The result from round 1 was taken as the result from a single LLM (GPT-4), while cases involving two or more rounds were classified as multi-agent scenarios.

We replicated their study to obtain the costs of the single LLM and the multi-agent system in their work, and divided the accuracy achieved by each by their respective costs, yielding their respective economic efficiencies. The entire estimation procedure followed rigorous engineering assumptions (see **Cost_Analysis.xlsx, sheet 6: Methodology**) and used the historical pricing of OpenAI GPT-4 as the benchmark for calculation. The specific method is as follows:

1. Parsing & Normalization

We wrote a parser to extract structured dialogues from each transcript. Dialogues were divided into separate discussion rounds based on numbered turns in the text (e.g., [Round 1]), and each speaker (AI_A to AI_E and Facilitator) was strictly identified. During preprocessing, we made auditable fixes to deterministic anomalies in the raw data, including standardizing the label "acilitator" to "Facilitator" and removing exact duplicate transcriptions.

2. Architectural Cost Scoping

To comprehensively cover computational costs, we constructed a two-tier cost estimation model:

Transcript Estimate (visible dialogue layer): This tier only accounts for the visible calls explicitly recorded in the original experimental transcripts. Specifically, each discussion round generated 5 AI agent response calls and 1 Facilitator summary/follow-up call.

Full Architecture Estimate (complete architecture layer): Although the Facilitator in the original experiments integrated all agent responses in a single call, in actual system deployment or under stricter cost ceiling estimations, the system often needs to perform consistency checks across multiple agent outputs. Specifically, after the five agents each provided their answers, the system would theoretically need to check for disagreement between any pair of answers. This pairwise logical judgment corresponds to C(5,2) = 10 pairwise comparisons. Therefore, we incorporated these 10 hidden "agree/disagree" comparison calls into the cost model for the full architecture, as a reasonable and auditable boundary estimate of the reasoning workload underlying the system. This assumption makes our cost estimation more conservative, covering the full computational footprint from surface-level interactions to underlying logical verification. Building on the visible layer, this tier includes the implicit reasoning overhead behind the multi-agent system. We assumed that the system performed internal consistency checks across the five agent answers, amounting to C(5,2) = 10 hidden pairwise comparison calls to simulate the

logical judgment of "agree/disagree."

3. Tokenization & Pricing

For each estimated API call, we counted the combined token length of the System Prompt and User Prompt as input cost, and the length of the generated response as output cost. Token counting was performed using the `tiktoken` library where available; when unavailable, we used `ceil(UTF-8 bytes / 4)` as an approximation. Notably, this estimation strictly excluded ChatML formatting overhead and additional costs from network retries, aiming to reflect the most raw infrastructure compute costs.

Guideline Evaluation

We assessed the reporting completeness of the included studies using the CHART guideline[27]. The CHART guideline was developed and published in 2025 by an international multidisciplinary team and is the first reporting guideline specifically designed for Chatbot Health Advice (CHA) studies. It consists of 39 items organized into 12 domains, covering title and abstract, introduction, methods (model identification, model details, prompt engineering, query strategy, performance evaluation, sample size, data analysis), results, discussion, and open science.

Two reviewers independently evaluated each included study. For each item, they assessed whether it was reported based on the full text, supplementary materials, and publicly available code repositories (if any). A three-category classification was used: "yes" if the item was clearly reported, "no" if it was not reported or the information was insufficient to determine, and "not applicable" if the item did not apply to the study design.

The results were summarized as counts and visualized as a heatmap to display the distribution of reporting completeness across items and studies.

Results

Study selection and characteristics

A total of 1281205 studies were initially identified from the databases by two independent reviewers, following a pre-specified search strategy and data collection protocol. Manual verification showed that the overall agreement rate between the LLM and human reviewers was 99.50%. The Cohen's kappa coefficient was 0.69 (95% CI: 0.67–0.71), indicating substantial agreement between AI and human judgments. We therefore adopted the LLM screening results (**Supplementary Material: Reliability Verification**). After preliminary screening, 587 studies underwent full-text review, and 49 studies (Including 46 original research articles[7,10,20,25,26,28-68] and 3 review articles[12-14]) ultimately met the eligibility criteria for inclusion in this synthesis. The specific reasons for exclusion at each stage were documented and are presented in the **Figure 1**.

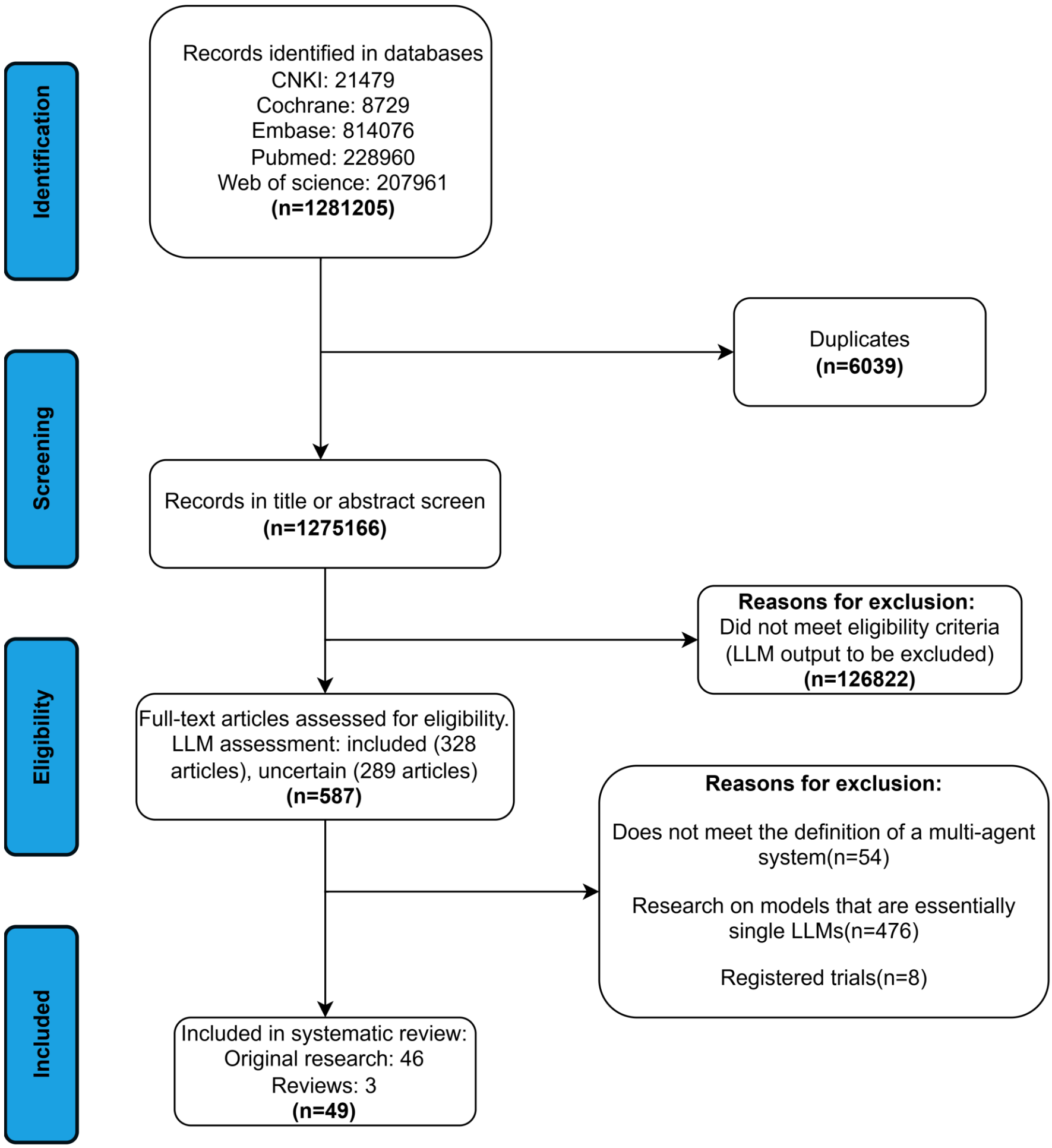


Figure 1: The flow diagram illustrating the process of literature identification, screening, eligibility assessment, and final inclusion. A total of 1,281,205 records were retrieved from five databases. After duplicate removal, title/abstract screening and full-text evaluation, 49 articles (46 original studies and 3 reviews) were included. CNKI: China National Knowledge Infrastructure.

A total of 49 articles were included in this study (detailed characteristics are provided in **Supplementary Table 3**). At the time of the search, the highest number of publications was 23 in 2025, followed by 23 in 2026, and 3 in 2024. Among these, 44 were peer-reviewed journal articles, with 2 preprints and 3 review articles. Regarding the type of multi-agent collaboration, among the original studies, 24 adopted single-round collaboration and 22 adopted multi-round collaboration.

In terms of the underlying model composition, the included studies showed considerable diversity. The GPT series was the most widely adopted model family, appearing in 29 studies across variants including GPT-3.5-turbo, GPT-4, GPT-4o, and GPT-o1-mini, reflecting the dominant position of OpenAI models in medical multi-agent research. The Qwen series (12 studies) and Llama series (12 studies) ranked next. Claude (10 studies) and DeepSeek (9 studies) were also commonly used.

In terms of task types, the included studies covered multiple medical application directions. Ten studies addressed medical knowledge question answering[10,25,26,29,32,33,51,52,63,66], nine focused on disease diagnosis[7,28,34,42,46,49,58,62,64], eight on research workflow automation[30,32,37,43,45,56,60], six on Alzheimer's disease (AD)[47,53,57,59,65,66], five on clinical text generation and summarization[31,39,40,50,64], and 11 fell into other categories[35,36,38,41,44,48,54,55,61,67,68].

Bibliometrics

Quarterly publication trend

From the third quarter of 2024 to the second quarter of 2026, the overall publication volume of studies on LLM-based multi-agent systems in medicine showed an increasing trend, with the line chart of total publications clearly reflecting the upward trajectory. A peak was observed in the first quarter of 2026, with 20 articles published in that quarter. In the initial stage (Q3–Q4 2024), research activity was minimal, with only 2 and 1 articles published in the two quarters, respectively, and nearly all contributions came from the United States, while most other countries had not yet initiated relevant studies.

Country distribution

After aggregating publications by institutional affiliation, marked differences were observed across countries. The China contributed the most publications, with 17 articles accounting for 34.7% of the total, followed by United States with 16 articles (32.7%). Together, the two countries accounted for over two-thirds (67.4%) of all included articles, making them the core contributing countries in this field. See **Figure 2-A** for details.

Journal distribution and impact (Figure 2-B)

The majority of publishing journals were in JCR Q1 (red bubbles), located in the high impact factor and high CiteScore range. Q2 (blue), Q3 (green), and Q4 (purple) journals were fewer in number, and a small number of unranked journals (gray) had relatively low impact overall. These findings suggest that studies in this field are predominantly published in high-quality Q1 medical journals, indicating a relatively high level of recognition. The journal with the highest number of articles was npj Digital Medicine, which published 5 articles and had the largest bubble size; it belongs to Q1 and has high impact factor and CiteScore. The second most productive journal was JCO Clinical Cancer Informatics, with 2 articles; it is in Q2 and has moderate impact. Most Q1 journals were concentrated in the range of IF 5–20 and CiteScore 5–25, with higher-impact journals tending to publish more articles, suggesting that prestigious journals are more willing to accept work on LLM-based multi-agent systems in medicine. A small number of lower-tier journals were clustered at the lower end of the axes and published few articles.

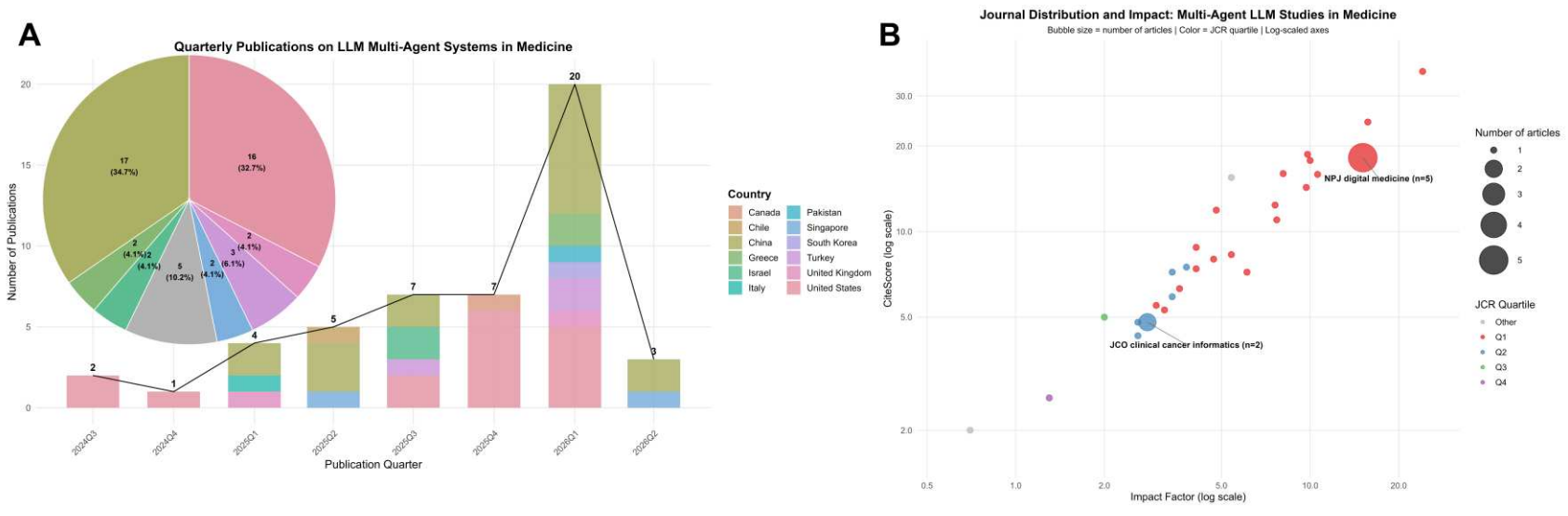


Figure 2: Publication trend, country distribution and journal characteristics of studies on LLM multi-agent systems in medicine. Panel A: The pie chart on the upper left illustrates the publication proportion of each country. The stacked bar chart below depicts quarterly publications from Q3 2024 to Q2 2026; different colors denote publications contributed by distinct countries, and the black line reflects the total quarterly publication volume. Panel B: Bubble plot showing journal distribution. The x-axis represents log-transformed journal impact factor, while the y-axis represents log-transformed CiteScore. Bubble size corresponds to the number of articles published in each journal, and colors classify journals according to JCR quartiles (red for Q1, blue for Q2, green for Q3, purple for Q4, gray for journals without quartile). The top two journals with the largest publication numbers are labeled with names and article counts.

Meta-analysis

Study characteristics

This meta-analysis included 30 comparisons from 8 original studies, with publication years ranging from 2024 to 2026. Sample sizes varied considerably across studies, ranging from 22 to 115722. By collaboration type, ‘Single-Round Multi-Agent Collaboration’ comprises two studies, whilst ‘Multi-Round Multi-Agent Collaboration’ comprises six studies.

Quality Assessment

Overall risk of bias: Among the 8 included studies, 6 (75.0%) were rated as low risk of bias, 2 (25.0%) as moderate risk, and none as serious risk. Of particular note, for Domain 6 (Bias in selection of the reported result), all 8 studies (100%) were rated as moderate risk. The primary reason was that none of the included studies had pre-registered a study protocol, leaving no documentation to verify pre-specified outcome measures. Although there was no clear evidence of selective outcome reporting, the possibility of potential publication bias could not be ruled out. The risk of bias assessment results are presented in **Figure 3**.

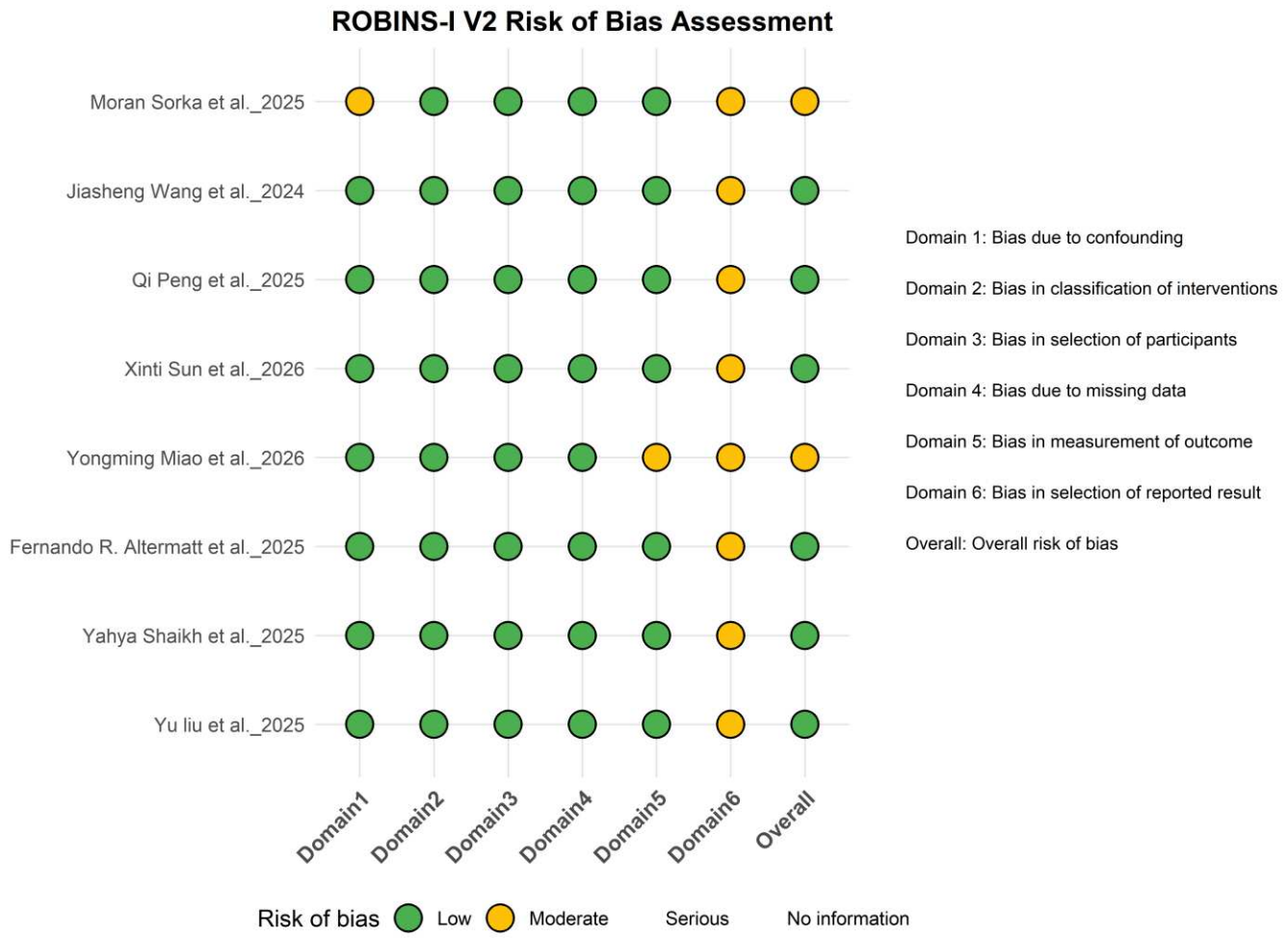


Figure 3: Risk of bias assessment of the 8 included non-randomized studies using the ROBINS-I V2 tool. Green circle indicates low risk of bias, yellow circle indicates moderate risk of bias. Domain 1: Bias due to confounding; Domain 2: Bias in classification of interventions; Domain 3: Bias in selection of participants; Domain 4: Bias due to missing data; Domain 5: Bias in measurement of outcome; Domain 6: Bias in selection of reported result; Overall: comprehensive risk of bias. No studies were rated as serious risk or no information.

It should be noted that the ROBINS-I tool was primarily designed for traditional non-randomized intervention studies, whereas the studies included in this review are mostly benchmark comparisons of AI models, algorithmic performance experiments, or retrospective dataset evaluations. These study designs differ from the original context for which ROBINS-I was intended, and their specific sources of bias may not be fully captured by the tool. Therefore, the ROBINS-I-based results presented above only reflect bias risk in terms of traditional methodological dimensions and should not be directly equated with acceptable overall quality of the evidence.

In this study, we also used a supplementary framework specific to AI comparative studies (see **Supplementary Table 4**) to systematically assess the above-mentioned biases, in order to provide a more comprehensive judgment of evidence quality. Regarding base model comparability, 5 studies (62.5%) ensured that the multi-agent system and the single model used the same base model and parameter settings. For prompt transparency, only 3 studies (37.5%) reported complete prompts, 4 (50%) did not, and 1 did not clearly specify. All 8 studies did not state whether the prompts were tuned on the test set, nor did they declare whether the test set questions were excluded from the model training data (only 1 study made such a declaration). In terms of resource parity, 5 studies (62.5%) ensured equal access to tools/RAG, and 4 (50%) reported equal token budgets. For randomness reporting, 3 studies (37.5%) reported random seeds, number of repetitions, and result variability; 2 did not report; and 3 reported partially. Regarding blinding, only 1 study reported blinded evaluators, while 5 did not report.

Inclusion in the Study and Overall Effects

This study included 8 studies with 30 comparisons. The results showed that multi-agent systems showed higher performance LLMs. The pooled OR was 2.22 (95% CI: 1.31–3.74), and the difference was statistically significant ($p < 0.001$) (**Figure 4-A**).

An Analysis of the Sources of Heterogeneity

There was significant statistical heterogeneity among the studies. The three-level model showed that the overall heterogeneity (I^2_{total} = 97.1%) consisted of between-studies heterogeneity (91.2%) and within-study heterogeneity (6.0%). The heterogeneity test (QE = 1197.01, df = 29, $p < 0.001$) indicated substantial heterogeneity across studies, likely due to differences in task types, model versions, or evaluation datasets used. Therefore, a random-effects model was appropriate.

After removing the four large-scale data arms from Qi Peng et al. (2025)[51], the heterogeneity metrics decreased substantially: overall I^2 (I^2_{total}) dropped from 97.1% in the original model to 31.4%. Between-study heterogeneity fell to 29.5%, and within-study heterogeneity to 1.9%, while the Q statistic decreased sharply from 1197.01 to 152.31. This indicates that nearly 70% of the heterogeneity in the original model was primarily driven by the large-sample data arms from Qi Peng et al. (**Figure 4-B**). The core statistical reason for this phenomenon is the extreme imbalance in weight allocation due to sample size. The study by Qi Peng et al. involved very large sample sizes. In a random-effects model using inverse-variance weighting, studies with large sample sizes have very small standard errors, thus receiving disproportionately large weights. Although the OR values of that study were not extremely discrepant from those of other studies, the subtle deviation in effect size was statistically amplified by the enormous weights, ultimately resulting in the extremely high heterogeneity of nearly 97.1% in the original model.

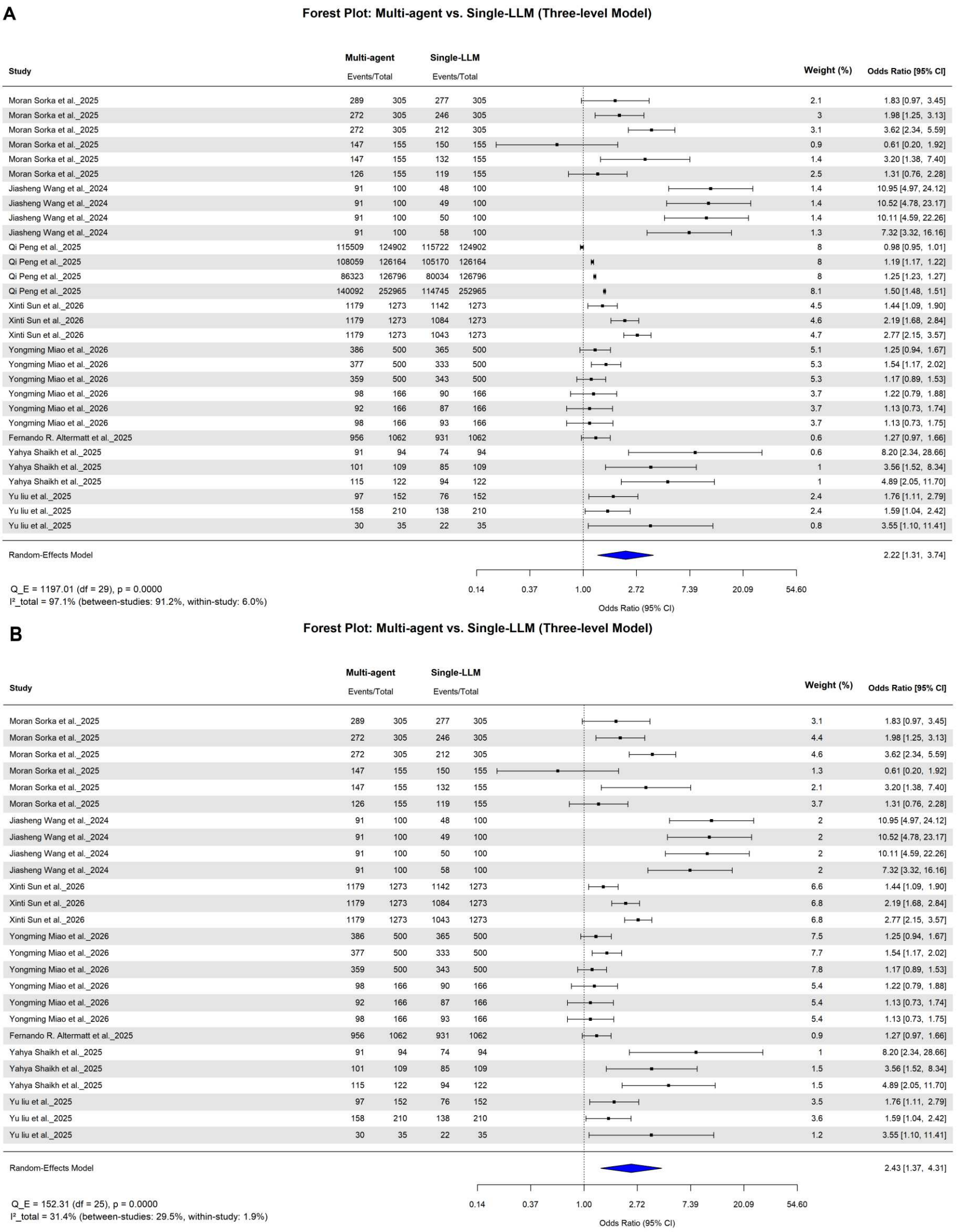


Figure 4: A three-level meta-analysis forest plot comparing multi-agent LLMs with single LLMs in terms of diagnostic and functional performance. (A) Main three-level random-effects analysis. (B) Sensitivity analysis after excluding one study. The diamond markers represent the pooled ORs and their 95% confidence intervals. An OR > 1 indicates that the multi-agent LLM system yields superior results.

Sensitivity Analysis

Leave-one-out sensitivity analysis: We refit the three-level model after sequentially excluding each study. After excluding any single study, the pooled OR ranged from 2.16 to 2.40, and all 95% CIs did not cross 1, which was highly consistent with the overall results (**Supplementary Figure 1**). In particular, after excluding the above-mentioned large-sample study with highly skewed weights, the refitted forest plot showed a pooled OR of 2.43 (95% CI: 1.37–4.31, $p < 0.001$, **Figure 4-B**), closely matching the pooled OR from the original three-level model. Moreover, the 95% CI remained entirely to the right of the null line of 1.0, and the core conclusion that multi-agent systems showed higher performance LLMs remained unchanged. This indicates that the overall meta-analysis conclusion was robust and not substantially altered by the exclusion of any single

study.

Correlation (ρ) sensitivity analysis: Given that the inclusion of multi-arm studies might introduce correlations among effect sizes within the same study, we set ρ to 0.2, 0.5, and 0.8, respectively, to correct the variance–covariance matrix (V matrix). The results showed pooled ORs of 2.19 (95% CI: 1.30–3.71), 2.15 (95% CI: 1.27–3.65), and 2.11 (95% CI: 1.26–3.55), respectively, with all p-values < 0.01. This indicates that even after adjusting for within-study correlation, the pooled results remained stable and significant.

Publication Bias Assessment

Funnel plots were used for qualitative assessment of publication bias, supplemented by Egger's linear regression test for quantitative verification. The funnel plot showed a roughly symmetric distribution (**Figure 5**), and Egger's test did not reveal significant publication bias (p = 0.4841).

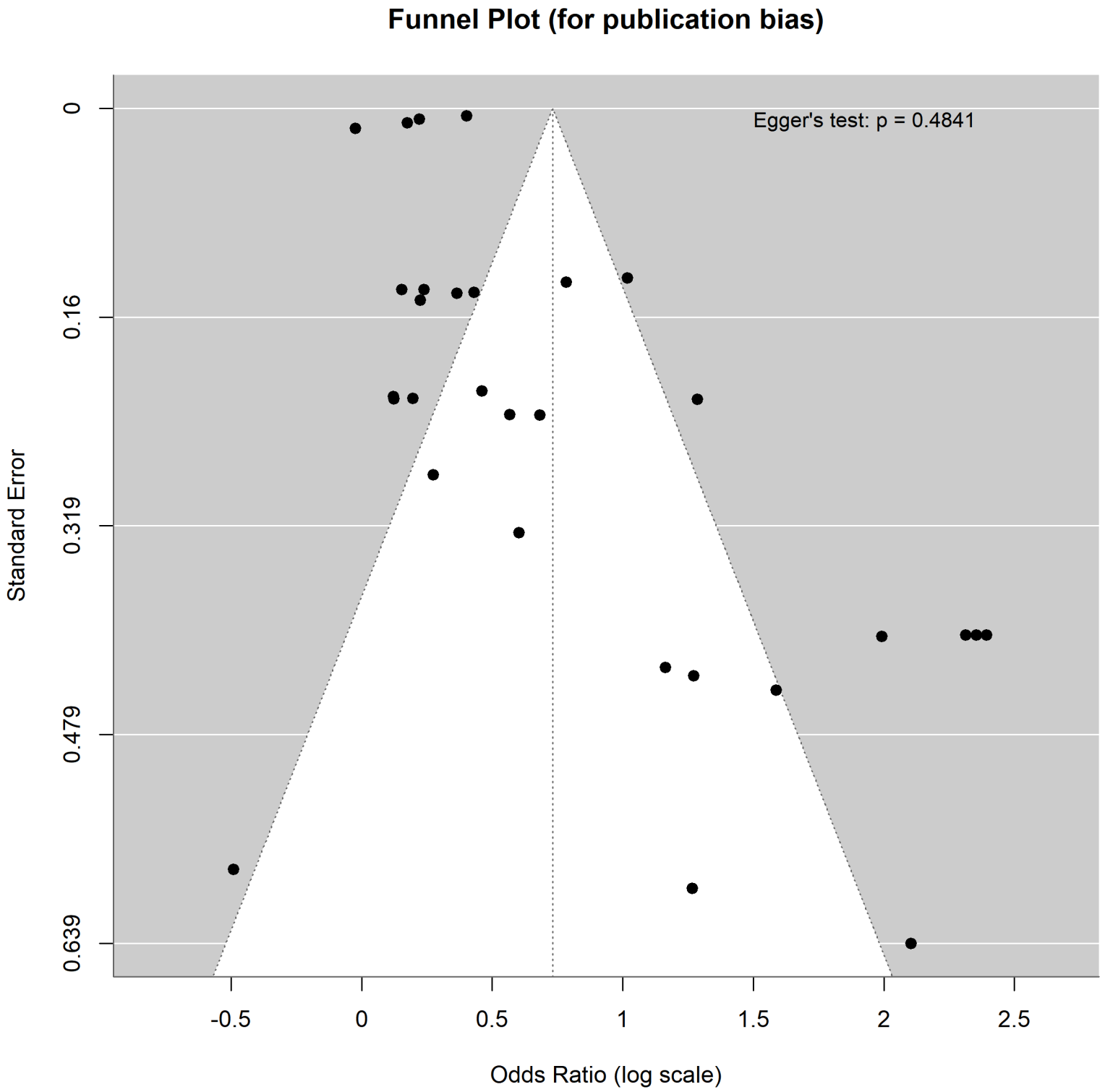


Figure 5: Funnel-plot assessment of publication bias for the pooled odds-ratio effect size comparing multi-agent and single-agent large-language-model performance. The horizontal axis represents log-transformed OR, and the vertical axis denotes the standard error of each individual study. The dashed triangular boundaries indicate the 95% confidence interval for the pooled effect. Egger's regression test yielded p = 0.4841, which suggested no statistically significant evidence of publication bias.

Paired Subgroup Sensitivity Analysis (Within-Study Pooling)

For the paired data from four different model configurations in the study by Jiasheng Wang et al.[32] (See Table 1 for details), the multi-agent system corrected errors made by the single model on 174 questions ($bpooled = 174$) and introduced new errors (i.e., changed correct answers to incorrect) on 15 questions ($cpooled = 15$). The pooled paired OR was 11.60 (95% CI: 6.85–19.65), and the exact McNemar test gave P < 0.001. This result indicates that, within the dataset of that study, the multi-agent system consistently showed a statistically significant net advantage across all tested configurations. As a single-center internal replication, this finding should be considered exploratory evidence, and its robustness requires further confirmation from additional independent studies.

Table 1: Results at the question level for multi-agent and single LLM responses in the research report

Study	Model	a	b	c	d
Jiasheng Wang et al.	Claude 3.7	47	44	2	7
	DeepSeek-R1	54	37	4	5
	Gemini-2.5-flash	45	46	5	4
	GPT-4o	44	47	4	5

Table 1: a, b, c and d represent the outcomes when a multi-agent system and a single LLM answer medical questions, respectively: both are correct; only the multi-agent system is correct; only the single model is correct; and both are incorrect.

Two additional question-answering studies did not provide 2×2 contingency table data for comparison with single models and were therefore not included in the meta-analysis; they are described below as a supplementary narrative.

Adeel Mukhtar et al. (2026)[29] developed a three-tier generation–critique–synthesis multi-agent conversational system for Parkinson's disease, with Qwen3-Medical-GRPO as the core model. The system accessed 80 medical knowledge bases through a RAG framework and incorporated knowledge graphs for personalized responses. On 50 clinical queries, the system achieved a clinical accuracy of 95%, with diagnostic suggestion scores of 4.8/5 and treatment recommendation scores of 4.6/5, and an average response time of 6.5 seconds. This system followed a single-round pipeline collaboration mode.

Cağatayumut Öğdü et al. (2025)[52] proposed a decision support system comprising five agents: clinical text clarification, symptom RAG analysis, evidence-based scanning, clinical data fusion, and adaptive optimization. The system used a BioBERT vector database and web-based evidence retrieval, triggering a feedback loop when confidence fell below a threshold ($\tau = 0.65$). On MedQA, PubMedQA, and MedBullets, the system achieved accuracies of 94%, 88%, and 84%, respectively. This study adopted a multi-round reflective evidence-integration mode.

Both studies demonstrated high absolute performance of multi-agent collaboration in MedQA. However, the absence of standard comparative data against single models prevented quantitative assessment of the relative benefit, highlighting the need for improved reporting standardization in this research area.

A Comparison of the Economic Benefits of Multi-Agent Systems and Single LLMs

Replication of the study by Yahya Shaikh et al.[26]:

The historical pricing for OpenAI GPT-4 was $30 per million input tokens and $60 per million output tokens (https://developers.openai.com/api/docs/models/gpt-4).

During preprocessing, corrections were made to two files:

In the transcript for STEP2-Question-31, a label marked as "acilitator" (missing the initial "F") was automatically standardized to "Facilitator".

In the transcript for STEP2-Question-96, the parser detected that the full transcript consisted of two identical halves in the same order, and automatically removed the duplicate second half to avoid double-counting (for details, see **Cost_Analysis.xlsx, sheet5: Data Quality**).

The total cost for GPT-4 (i.e., Single-round) was $229.99, and the total cost for the multi-agent system (i.e., Overall = Single-round + Multi-round) was $411.90 (for details, see **Cost_Analysis.xlsx, sheet1: Dashboard**).

Economic benefits

GPT-4 achieved an average accuracy of 78.00% on the questions in this task, while the multi-agent system achieved 94.67%. Thus, the economic benefit of GPT-4 was $0.34 per percentage point of accuracy, and that of the multi-agent system was $0.23 per percentage point of accuracy.

The economic benefits results for the two studies are summarized in **Table 2**.

Table 2: Summary of economic benefits for the two studies.

Study	Single LLM	Single LLM total cost	Multi-agent total cost	Single LLM accuracy	Multi-agent accuracy	Single LLM economic benefits	Multi-agent economic benefits
Yahya Shaikh et al.[26]	GPT-4	$229.99	$411.90	78.00%	94.67%	0.34	0.23
Xinti Sun et al.[25]	GPT-o1	$9.7	$25.5	85.15%	92.62%	0.11	0.28
Xinti Sun et al.[25]	Qwen-QwQ	$3.6	$25.5	81.93%	92.62%	0.04	0.28
Xinti Sun et al.[25]	DeepSeek-R1	$5.7	$25.5	89.71%	92.62%	0.06	0.28

Table 2: The unit of economic benefit is: $ per % accuracy

The study by Yahya Shaikh et al. found that to achieve the same level of accuracy, the single LLM incurred higher costs than the multi-agent system, whereas the results from Xinti Sun et al. showed the opposite. This suggests that cost is related to the multi-agent architecture, and that exploring economically efficient multi-agent architectures is feasible.

Multi-agent Applications for Other Tasks

Disease diagnosis

Based on nine original studies (**Supplementary Table 5**), multi-LLM frameworks achieved accuracy exceeding 90% in

moderate-complexity diagnostic tasks (3–10 classes). In extremely high-complexity scenarios such as rare diseases (>10 classes or >300 diseases), Top-1 accuracy ranged from approximately 34% to 44%, while Top-3 recall reached over 55%, demonstrating good value for differential diagnosis. Compared with single models or conventional clinical pathways, multi-LLM systems showed higher diagnostic concordance and significantly shorter decision-making time (the complete process taking only a few minutes), and some studies have achieved pathological prediction performance surpassing that of human experts (e.g., 4R-tau subtyping in primary progressive aphasia). However, this paradigm still faces common limitations, including high computational costs, lack of multicenter prospective validation, and insufficient integration with real-world workflows. Overall, current evidence supports multi-LLM systems as scalable and time-efficient assistive diagnostic tools, with further optimization focusing on lightweight deployment, human-AI collaborative interaction, and the development of continuous learning mechanisms.

To validate the effectiveness of multi-agent collaboration, most studies included head-to-head comparisons against a single LLM using the same base model. The results consistently showed that multi-agent frameworks outperformed single models in diagnostic accuracy, alignment with clinical decisions, and robustness on complex tasks, with the advantage widening as task complexity increased. Using the data from **Supplementary Table 6** as an example: in digestive disease diagnosis, DeepGut improved the accuracy of a single-modality LLM from 52.9% to 97.8%; in a five-class classification task for intestinal obstruction and volvulus, a multidisciplinary AI system outperformed ChatGPT-5.0 by 7.2 percentage points in accuracy and 7.3 percentage points in surgical decision alignment; in the extremely low-prevalence scenario of 302 rare diseases, the MAC framework raised GPT-4's Top-1 accuracy from 19.7% to 34.1%, representing a relative improvement of over 70%. In a three-class AD classification task, AD-Multi-Agent achieved a weighted F1 score of 88%, significantly outperforming the best single-dimensional agent (78%). Collectively, these comparisons confirm that incorporating mechanisms such as role specialization, debate-based validation, and experience retrieval is a critical pathway to unlocking the clinical diagnostic potential of LLMs.

Research workflow automation

LLM-based multi-agent systems have made progress in automating biomedical research workflows, covering tasks ranging from literature screening and data extraction to experimental protocol generation, data analysis, and even interpretation of complex clinical cases.

In literature screening and data extraction, researchers aim to address the inefficiency and error-proneness of manual screening of large volumes of literature. Pan et al.[45] adopted a three-party collaborative debate framework involving initial review, commentary, and arbitration to screen literature on liver cancer treatment. Their approach accurately identified truly relevant articles (weighted recall = 0.90) and achieved high agreement with human expert judgments (PABAK = 0.91), suggesting that multi-model collaboration effectively reduces missed articles. However, the weighted precision was only 0.15, meaning that a large number of irrelevant articles were incorrectly flagged as included, necessitating substantial manual cleanup of the screening results.

Akinseloyin et al.[56] systematically compared three strategies—majority voting, multi-agent debate, and LLM adjudication—and found that simple majority voting achieved the highest average precision (0.341 and 0.462 for diagnostic technology assessment and intervention reviews, respectively) and work savings (WSS@95% reached 0.680 and 0.606), with costs less than 1/14 of the adjudication method. The key to its success lay in preserving heterogeneity across models—debate, while improving the performance of weaker models, tended to homogenize opinions and reduce diversity, thereby weakening the ensemble gain.

Khan et al.[60] simulated a dual-reviewer process by having GPT-4-turbo and Claude-3-Opus independently extract data. They found that consensus responses between the two models achieved an accuracy of 0.94 with a hallucination rate of only 0.25%. When discrepant responses underwent cross-review, 51% reached new consensus, with accuracy rising from 0.4–0.5 to 0.76; however, nearly half of the disagreements remained unresolved, and variables requiring implicit judgment performed poorly.

Multi-agent collaboration significantly improves recall and consistency, but the precision bottleneck and limited capacity for complex judgments remain major obstacles in automated literature processing. Designing mechanisms that maintain diversity while efficiently resolving disagreements will be key to future breakthroughs.

In the domain of biomedical data analysis and experimental protocol generation, Luo et al.[30] developed an end-to-end system named BioResearcher, which automated the entire workflow from research objectives to executable code through a modular multi-agent pipeline comprising search, literature processing, protocol design, and programming. Across eight unsolved research goals, the average execution success rate reached 63.07%, and the quality of generated experimental protocols was 22.0% higher than that of typical agent systems. Among the failed tasks, 67.19% involved minor errors that did not require substantial manual intervention. The primary failure modes were data transformation issues (46.4%) and code errors (21.4%), with output truncation due to lengthy input texts also warranting attention.

Similarly, Wang et al.[32] designed a role-playing multi-agent system (administrator, planner, data retrieval and cleaning agent, data scientist, reviewer) for clinical data analysis. The system successfully reproduced the primary outcomes of 53.3% of published studies, significantly outperforming the single ChatGPT-4o model at 35.0%, and no variable or outcome hallucinations were observed. Failures were likewise dominated by data transformation (46.4%) and incorrect statistical methods (17.8%), whereas the single LLM, in its eagerness to produce answers, more frequently switched to inappropriate methods (38.4%).

Liveratos et al.[43] integrated a dedicated LLM-based retriever, a multi-agent runtime framework (first layer with multiple open-source LLMs, second and third layers with GPT-4o for aggregation and validation), and automated code generation to automate network meta-analysis, compressing the original months-long workflow to within one week. Although the system

performed efficiently in specific, standardized scenarios, its reliance on opaque closed-source models, limited adaptability to complex and heterogeneous data, and lack of automatic within-model adjustment for heterogeneity indicate that it remains far from reliably replacing human-conducted meta-analyses in a broad context.
Taken together, these studies demonstrate that multi-gent systems, through specialized role allocation and iterative feedback, can effectively enhance the automation and robustness of complex tasks. However, they remain susceptible to fundamental errors such as data transformation issues, and their success still heavily depends on carefully designed prompts and relatively standardized data formats, suggesting that fully autonomous operation without human intervention is not yet within reach.

In the domain of complex clinical data interpretation and error detection, Hedaya et al.[59] proposed a novel iterative dual-AI cross-validation method. In a contradictory case where imaging suggested "brain volume increase" but cognitive tests showed decline, Claude Sonnet 4 and Perplexity initially produced markedly divergent judgments after independent analysis (probability estimates differing by 45–60%). Through 5 to 7 rounds of mutual review, questioning, and revision, they eventually converged to a consensus (discrepancy < 10%). During this process, the AI autonomously identified critical errors—a 3.5% increase in total intracranial volume (physiologically impossible, indicating a measurement artifact) and an 11-month time gap between cognitive testing and MRI—and constrained the estimated effect of hyperbaric oxygen therapy to a maximum of 1–2% volume change through literature retrieval. This simulated peer-review mechanism effectively identified hallucinations and logical gaps that single models tend to overlook. However, a potential risk remains: if both models share the same systematic bias, they may converge to a false consensus. Furthermore, the entire process still required human orchestration and had not achieved full automation. This suggests that multi-agent collaboration can serve not only for task execution but also as a powerful tool for quality assurance and error detection. The integration of this approach with the aforementioned data extraction and protocol generation modules may represent the ultimate pathway toward a credible and deployable automated research system. Details of each study are provided in **Supplementary Table 7**.

Applied to Alzheimer's disease

Simulating clinical team-based diagnostic reasoning represents the prevailing paradigm in current research. Li et al. [47]proposed the CARE-AD framework, which comprises six independent LLM agents emulating data extraction, primary care screening, neurological evaluation, psychiatric assessment, geriatric assessment, and clinical psychological analysis, with a final AD specialist agent synthesizing the findings. By analyzing longitudinal unstructured EHR texts, CARE-AD predicted AD risk up to 10 years before formal diagnosis with an accuracy of 0.53, significantly outperforming single-model chain-of-thought (0.35) and self-consistency (0.43) baselines. This study was the first to demonstrate the unique advantage of multi-agent collaboration in extracting long-term risk signals from unstructured clinical narratives. In a parallel effort, Tian et al. (2026) developed a fully autonomous five-agent workflow for detecting cognitive impairment from clinical notes. Unlike CARE-AD, their system required no human intervention, with agents autonomously refining prompts to achieve an F1 score of 0.93 on a balanced dataset; however, sensitivity dropped to 0.62 under real-world prevalence (33%), highlighting the challenge of training-deployment data distribution shifts on multi-agent generalization.

Extending to multimodal diagnosis and therapeutic discovery represents another important direction in multi-agent AD research. Zhang et al. (2026)[65] proposed a self-learning and critique framework comprising 14 agents, with 9 unimodal agents and 5 higher-level diagnostic agents working in tandem. The framework achieved an F1 score of 88% for CN/MCI/AD three-class classification using only role-based prompting and experience retrieval, without end-to-end supervised training on the target dataset, and provided modality-level attribution explanations. This work demonstrates that multi-agent architectures can effectively integrate heterogeneous multimodal data while maintaining high diagnostic performance in the absence of large-scale labeled samples. In the domain of drug discovery, the Coated-LLM framework (Xu et al., 2025[57]) employed a three-tier "researcher–reviewer–arbitrator" agent structure to predict combination therapy efficacy for AD under extreme data scarcity. The framework achieved a prediction accuracy of 74%, substantially exceeding the 52% of traditional network-based methods, and further validated the biological effects of one predicted drug combination (m266 + gypenoside) through in vitro experiments, completing a closed loop from computational prediction to wet-lab validation.

Knowledge augmentation and error self-checking mechanisms have become key technical pathways for improving the reliability of multi-agent systems. Liu et al. (2026)[66] proposed the KG-MRL framework, which combines knowledge graph retrieval with multi-role evaluation. Through a three-round iterative refinement involving critic, improver, and arbiter agents, it achieved an average accuracy of 74.8% on AD question-answering tasks, approximately 13 percentage points higher than the baseline GPT-3.5. This indicates that the synergy between structured external knowledge and multi-agent self-critique can effectively mitigate LLM factual hallucinations. Furthermore, Hedaya et al. (2026)[59] reported an instructive case study in which two independent AI systems, Claude and Perplexity, engaged in 5 to 7 rounds of iterative cross-examination. Their interpretations of the same complex neuroimaging case converged from an initial 45–60 percentage point divergence to less than 10 percentage points, and they autonomously identified measurement artifacts and literature evidence overlooked by human experts. This work is the first to demonstrate, in a case-study format, how "AI peer review" can achieve convergent validity. Although its statistical generalizability is limited, it provides an important proof of concept for applying multi-agent systems to clinical error detection.

Although existing studies have demonstrated the promising potential of multi-agent LLMs in the AD domain, this direction remains in its early exploratory stage and faces several common challenges. First, data bias and generalizability are widespread concerns—studies have been based on the United States Veterans Health Administration cohort (CARE-AD, Li et al., 2025[47]) or AD Neuroimaging Initiative participants (Zhang et al., 2026[65]), with relatively homogeneous demographic characteristics (Tian et al.[53], 2026; Hedaya, 2026[59]). Their performance in real-world, multicenter, diverse populations

remains to be validated. Second, computational cost and inference latency cannot be ignored: multi-round agent interactions substantially increase token consumption and response time, potentially limiting deployment in real-time clinical decision-making (Xu et al., 2025[57]; Tian et al., 2026[53]; Liu et al., 2026[66]). Third, although some frameworks present the reasoning process or key evidence underlying their diagnostic conclusions (Li et al., 2025[47]; Zhang et al., 2026[65]), whether these "explanations" are merely post-hoc plausible justifications or genuinely reflect the models' internal decision-making logic currently lacks systematic evaluation methods. Finally, there is no unified evaluation benchmark or standardized comparison framework, and studies differ considerably in task definitions, data splits, and baseline selections, making direct cross-study comparisons difficult.

In summary, multi-agent LLMs offer a computational framework for clinical decision-making in AD that integrates specialized role allocation, knowledge augmentation, and self-correction capabilities. From CARE-AD, which simulates multidisciplinary consultations, to agentic workflows with self-optimized prompts, and KG-MRL combining knowledge graphs with multi-role evaluation, researchers are progressively shifting from single large models toward structured agent collaboration. Future work should validate the robustness of existing methods on larger, more diverse real-world datasets, develop more efficient and lightweight agent communication mechanisms, and establish standardized evaluation systems to advance this technology from proof-of-concept toward genuine clinical decision support tools. (See **Supplementary Table 8** for study details.)

Clinical text generation and summarization

The core task of clinical text generation and summarization studies is to transform unstructured raw medical information into high-quality structured text. The unique value of multi-agent systems in this context lies in their ability to achieve collaborative control through role division and iterative feedback mechanisms, addressing dimensions such as clarity, completeness, and factual accuracy that single models struggle to manage alone.

In diagnostic report generation, Chang Cai et al.[31] developed a generation-revision dual-agent system, in which a dedicated revision agent iteratively optimized initial reports against six criteria including balance and clarity. This improved report clarity from 3.0 to 4.89 and semantic analysis from 3.25 to 4.55, though performance on negative sentiment analysis remained insufficient, and the final reports still required manual insertion of figures, indicating that full automation had not yet been achieved. Alon Gorenstein et al.[39] decomposed electromyography report generation into sequential expert modules for data validation, knowledge retrieval, and comprehensive reporting, achieving a detection accuracy far exceeding that of a single model (92.2% vs 62.6%). However, the system tended to hallucinate or omit secondary diagnoses when encountering rare syndromes, revealing a strong dependence on the completeness of the knowledge base. Jifan Gao et al.[40] used specialized agents to zero-shot convert non-textual modalities (e.g., imaging, laboratory results) into textual summaries, which were then fused with clinical notes for prediction. They achieved a macro F1 of approximately 0.85 in trauma severity stratification, with stable performance across racial and sex subgroups. However, ablation experiments revealed an important boundary: when the added modality was not strongly associated with the prediction target, no performance gain was observed.

Turning to evaluation, Yuhao Chen et al.[50] explored using LLMs as judges. They found that GPT-4 achieved 83.06% agreement with at least one medical expert, but the four experts themselves had a complete agreement rate of only 19.2%. This contradiction directly questions the reliability of treating human experts as the gold standard for evaluation and suggests that LLM-based evaluators are better positioned as adjuncts rather than arbiters. In radiology reporting, Jia Li et al.[64] explicitly induced empirical rules from historical reports and combined them with sentence-level interpretable analysis, improving CheXpert F1 by 14.15%. Nevertheless, the generated impression quality still showed a ROUGE gap of over 50% compared with human experts, and the lengthy explanations added to clinicians' reading burden.

Overall, multi-agent systems have made considerable progress in improving the surface quality and structural organization of clinical text through modular design and iterative refinement. However, deeper challenges—including hallucination control, generalization to rare scenarios, and the inherent inconsistency of human standards—remain core barriers to moving from assisted writing to autonomous reporting. (See **Supplementary Table 9** for study details.)

Others

The clinical tasks addressed by the included studies varied considerably across a wide spectrum. We constructed a summary table of study characteristics (**Supplementary Table 10**) to qualitatively synthesize the applications of LLM-based multi-agent systems in other medical domains. This table comprises 11 representative studies, with each row corresponding to one independent study. Through this structured and information-dense presentation, we aim to provide a reference evidence map for future systematic reviews or methodological classifications, and to inform researchers in selecting appropriate collaborative architectures and validation strategies when designing medical multi-agent systems.

Guideline Evaluation

The results of the evaluation of individual items across the 46 original studies, together with their heatmap distribution, are shown in **Figure 6**. For items with relatively adequate reporting, clinical context (2a, 87.0%), study objectives (2b, 80.4%), statistical methods (9a, 80.4%), performance metrics (9ai, 93.5%), performance results (10a, 89.1%), and the discussion section (11a–c, 97.8%) all showed high reporting rates, indicating that the field has reached a basic consensus on study positioning and performance reporting. However, systematic reporting gaps exist in core methodological elements. Model name and version (3a, 41.3%) and open-source status (3b, 26.1%) were inadequately reported, with more than half of the studies failing to precisely specify the model version under evaluation. Prompt evolution (5a, 20%), full prompts (5b, 37.8%),

and prompt sources (5ai, 8.9%) were reported at extremely low rates—prompts constitute the sole input variable in CHA research, and failure to disclose prompt details renders a study irreproducible, equivalent to failing to report the dosing regimen in a clinical trial. Query date and location (6b, 0%), independent session use (6c, 0%), and full output provision (6d, 7.3%) were also severely underreported; the inherent randomness of LLM outputs requires researchers to report the temporal and environmental conditions of queries; otherwise, findings cannot be independently verified. Assessor blinding (7biii, 18.4%) and sample size justification (6.5%) also showed low reporting rates, affecting objectivity and statistical robustness. In terms of open science and ethical compliance, protocol preregistration (12d, 2.2%) was almost entirely absent, while ethical approval (12c, 34.9%) and data privacy protection (12ci, 27.3%%) also require substantial improvement, suggesting that this field remains in an early stage of methodological development.These deficiencies indicate that if CHA research is to accumulate high-level evidence for clinical decision-making, urgent efforts are needed to establish systematic standards for prompt transparency, query environment standardization, and study preregistration.

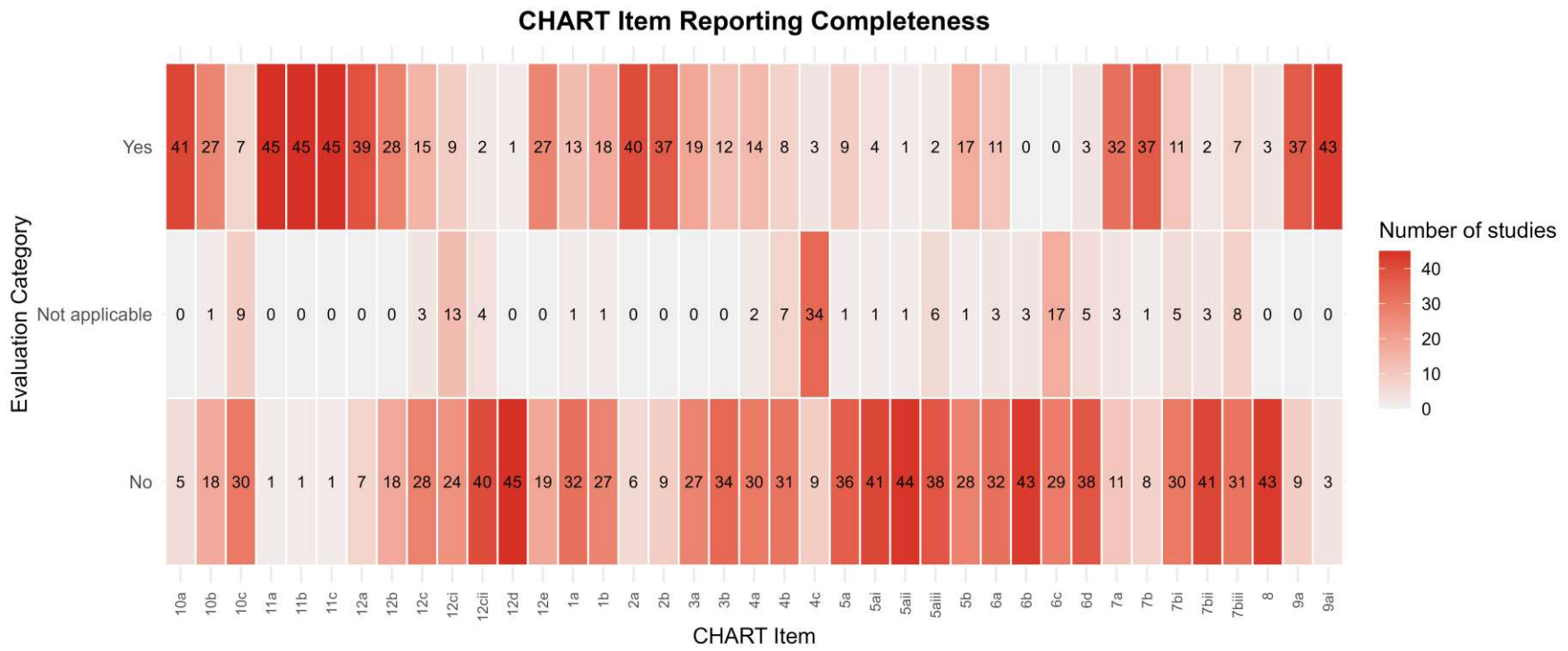


Figure 6: Heatmap showing reporting completeness of all items in the CHART checklist. The horizontal axis represents individual CHART items, and the vertical axis indicates three evaluation categories: reported adequately (Yes), not applicable, and not reported (No). Color intensity corresponds to the number of included studies, with darker red indicating more compliant studies. The heatmap intuitively displays heterogeneous reporting quality across different domains, with sufficient reporting in clinical context, outcomes and discussion, while prominent underreporting exists in model information, prompt details, experimental conditions, trial preregistration and ethics.

Descriptive resource and cost profile

Cost analysis of multi-LLM systems in medical applications needs to be considered from two dimensions: compared with human effort, they offer significant advantages in time and cost; compared with single models, they incur additional computational overhead (see **Table 3**). For instance, MetaMind compressed the network meta-analysis process from months to within one week while maintaining statistical rigor. A multi-LLM collaborative literature screening system reduced workload by 68%, with economic costs only 1/186 of that of human effort[56]. CASSIA completed single-cell annotation in 30 seconds per case at a cost of $0.02–0.03[57], whereas manual annotation takes several minutes and relies on specialized training. However, the multi-round debate, cross-critique, and consensus mechanisms of multi-LLM systems increase resource consumption. The MCC framework consumed an additional 4,814 tokens per case compared with single models in the MedQA task[25], with 19.5 seconds more time and approximately $0.005 additional cost per case. The MAC framework cost $0.12–0.17 per case, higher than single GPT-4. Cross-critique in dual-LLM collaborative data extraction added approximately $2.06 per article[60]. Nevertheless, moderate cost increases can yield higher reliability: cross-critique reduced hallucination rates from 2.5% to 0.25%[60], dual-LLM collaboration improved data extraction accuracy from 89% to 94%[60] , and AI-SC achieved 98.2% accuracy in contraindication identification for liver transplant evaluation[54]. Future strategies to further optimize costs include model lightweighting (e.g., LLaMA-3.2-90B reducing cost to $0.003 per annotation), task-selective activation of debate, on-premise deployment, and batch processing. In summary, some studies reported substantial time savings relative to human workflows, whereas multi-agent approaches generally incurred greater computational costs than single-model baselines.

Table 3. Cost-benefit analysis of multi-agent systems across representative medical applications.

Citation	Study_id	Application scenario	Time cost (per case)	Economic cost	vs. single model	vs. human	Key performance
33	Liveratos et al.	Automated network meta-analysis	<1 week (full process)	Not specified	Higher	Saves months	Accuracy 92.6%
35	Pan et al.	Automated literature screening for meta-analysis	36 min/dataset	$1.29/dataset	Higher	1/186 of human cost	Accuracy 96%
17	Wan et al.	Assisted diagnosis of digestive diseases	4–12× that of single model	Higher (more tokens)	+4–12× time	Better than single modality	Diagnostic accuracy 97.8%
44	Sun et al.	MedQA	67 sec	$0.020/case	$0.01	Much lower than human	Accuracy 92.6%
47	Akinseloyin et al.	Literature screening by titles/abstracts	36 min/dataset	$1.29/dataset	Lower for some strategies	1/186 of human cost	WSS@95% ≈ 68%
7	Chen et al.	Rare disease diagnosis	—	$0.12–0.17/case	+$0.02–0.05	Much lower than experts	Diagnostic accuracy 34% → 54%
26	Xie et al.	Cell type annotation	30 sec/cell	$0.02–0.03/cell	$0.01	Human takes minutes	Accuracy 92%
48	Xu et al.	Drug combination prediction for Alzheimer's disease	4 min	$0.95/case	$0.42	Reduces experimental cost	Precision 82%
51	Khan et al.	Automated data extraction for systematic reviews	—	$3.40/article	+$2.06 (cross-critique)	Saves weeks	Accuracy 94%
45	Hasjim et al.	Liver transplant candidate evaluation	—	Not specified (API limit)	Higher	Improves consistency	Accuracy 98.2%

Table 3: All cost and time metrics are as reported in the original studies.A dash indicates the metric was not reported. Costs should be interpreted alongside performance gains, as increased cost may be offset by improved accuracy or reliability. For WSS@95%, values represent the proportion of manual workload saved while maintaining a 95% recall rate. All monetary values are in US dollars. Abbreviations: vs.: versus; WSS@95%: Work Saved over Sampling at 95% recall; API: Application Programming Interface; MedQA: Medical Question Answering; LLM: Large Language Model; AI: Artificial Intelligence.

Discussion

This study provides a systematic overview of the landscape of multi-agent collaboration research in medicine through bibliometric analysis, meta-analysis, CHART guideline evaluation, and cost analysis. We made the following contributions: we introduced a three-level meta-analysis model to address the non-independence of multiple comparison pairs within the same study, designed a sensitivity analysis strategy with variance–covariance correction for missing paired data, and conducted paired subgroup validation. Together, these formed a reusable statistical framework that may serve as a methodological reference for future similar studies. Our findings show that the number of publications in this field has grown rapidly, with most articles published in JCR Q1 journals, indicating strong academic interest. The meta-analysis confirmed that multi-agent systems showed higher performance LLMs in accuracy on MedQA tasks ($p < 0.001$). The resource and cost profiles showed that multi-agent systems had substantial advantages over manual approaches in both time and cost, but the direction of economic efficiency relative to single LLMs varied by architecture, suggesting that exploring cost-effective multi-agent architectures is a feasible direction. However, the CHART guideline evaluation also revealed systematic deficiencies in methodological transparency and reproducibility in this field—reporting rates for key items such as prompt engineering, query environment, and preregistration were generally low. This indicates that the current level of evidence is insufficient to support clinical translation, and that more standardized reporting and validation systems are urgently needed in future research.

This study has several limitations. First, our search strategy was constructed around terms related to LLMs and specific model names, combined with multi-agent related terms and definitions for screening. However, there is currently no unified terminology for multi-agent systems across studies, and some relevant studies may have been missed due to inconsistent naming conventions. Second, the studies included in this review were predominantly in English and Chinese, potentially missing relevant studies published in other languages, constituting a language bias. Third, we used the ROBINS-I tool for risk of bias assessment, which was primarily designed for traditional non-randomized intervention studies, whereas the bias sources in AI model benchmark comparison experiments may not be fully captured by this tool. Although we supplemented with an AI-specific evaluation framework, this framework has not undergone formal validity testing. Fourth, most primary studies reported only marginal totals without providing question-level paired 2×2 table data, preventing precise calculation of paired effect sizes. Although we conducted sensitivity analyses using variance–covariance correction, this is not an ideal solution. Fifth, the CHART assessment indicated systematic deficiencies in methodological reporting, which limits the overall level of evidence. Sixth, this study reflects the state of the evidence as of mid-2026; as research evolves rapidly, effect sizes may revert to their original levels.

The temporal trend analysis indicates that multi-agent collaboration has emerged as a growing research focus in medical artificial intelligence, and studies in this field have gained recognition from high-quality academic journals. However, the

country-level distribution is highly concentrated in China and the United States, which together account for 67.4% of all publications, suggesting a substantial geographic imbalance in research resources. This pattern is not incidental but is closely tied to the economic scale and innovation capacity of the two nations. Prior research has demonstrated a significant correlation between medical AI research output and national economic and innovation indicators[69]. As the world's two largest economies and leading hubs for technological innovation, China and the United States possess considerable advantages in AI infrastructure, research and development investment, and talent pool—resources that provide the necessary foundation for conducting large-scale medical multi-agent research[70].

Nevertheless, this highly concentrated research landscape gives rise to concerns.

Previous systematic reviews have shown that 95.1% of medical AI patient cohorts originate from high-income or upper-middle-income countries, while low-income countries are entirely absent. In terms of racial composition, White and Asian patients together account for 91.7%[71]. The direct consequence of this data bias is that AI models can outperform humans in the source populations (accuracy 82.9% vs 77.3%), but when deployed to settings with different demographic characteristics, this advantage disappears (74.1% vs 76.3%). Furthermore, the research focus has been concentrated on disease spectra prevalent in high-income countries (e.g., cardiovascular diseases, cancer), while infectious diseases, parasitic diseases, and maternal health issues—which are more common in low-income regions—have received insufficient attention in mainstream AI research[72]. The proliferation of LLMs may further exacerbate global inequalities in healthcare resources[70]. Therefore, the external validity of current conclusions is largely limited to the healthcare systems of China and the United States, and their applicability to other regions requires validation with multicenter, multi-regional data. Future research urgently needs to establish more regionally representative datasets and international collaborative networks.

The meta-analysis demonstrated that multi-agent systems showed higher performance LLMs in MedQA tasks. However, this statistically significant difference needs to be interpreted cautiously in the broader clinical context.

There is a substantial gap between benchmark performance and clinical competence. Current studies on MedQA have relied almost exclusively on standardized benchmarks such as MedQA and PubMedQA[18]. These benchmarks typically adopt a multiple-choice format and primarily assess the model's ability to recall medical factual knowledge. However, multiple studies have shown that strong performance on standardized medical examinations does not directly translate to reliable clinical practice capability[73]. Kim et al. systematically analyzed 85 LLMs across 702 clinical evaluations and found that benchmark scores showed only a moderate correlation with real-world clinical performance (Spearman's rho = 0.59), and that mainstream benchmarks such as MedQA failed to capture core clinical competencies including patient communication, longitudinal care, and clinical information extraction[18]. A systematic review of 39 medical LLM benchmarks further quantified this knowledge–practice gap: models achieved 84%–90% accuracy on knowledge-based benchmarks, but this dropped to 45%–69% on practice-oriented evaluations and to only 40%–50% on safety assessments[73]. This systematic discrepancy suggests that effect sizes observed on multiple-choice benchmarks—whether from single models or multi-agent systems—cannot be directly extrapolated to clinical practice.

The pathway from bench to bedside presents multiple hurdles between positive benchmark results and actual clinical adoption. In an AJR expert review, Bhayana et al. pointed out that LLMs are not yet autonomous entities capable of replacing expert decision-making, and clinicians remain responsible for the outputs; patient-facing medical AI chatbots require additional safety guardrails[74]. Prospective clinical validation studies of multi-agent systems remain scarce—a notable gap from the high-level evidence expected by evidence-based medicine, such as prospective randomized controlled trials. Accordingly, the finding from our meta-analysis—that multi-agent systems significantly outperform single models in MedQA—should currently be interpreted as evidence of efficacy under controlled laboratory conditions rather than evidence of effectiveness in real-world settings. The actual clinical benefit of multi-agent systems in routine practice still awaits verification through prospective, multicenter clinical trials.

The core contribution of this meta-analysis is the demonstration that multi-agent collaborative systems achieve statistically significant performance gains over single LLMs in standardized MedQA tasks. However, given the systematic gap between benchmark performance and real-world clinical capability, the complexity of authentic clinical data, and the scarcity of prospective validation, this finding should be interpreted as evidence of superiority under controlled conditions rather than as a basis for clinical deployment. Future research should evaluate multi-agent systems within assessment frameworks that more closely resemble real clinical practice and conduct prospective clinical trials to generate higher-level evidence.

The findings from the meta-analysis of question-answering tasks are consistent with qualitative evidence from disease diagnosis, research workflow automation, Alzheimer's disease, and clinical text generation. In disease diagnosis, multi-agent frameworks improved GPT-4's Top-1 accuracy from 19.7% to 34.1% in high-complexity scenarios such as rare diseases, representing a relative increase of over 70%. In research automation, multi-agent systems have reduced the timeline for network meta-analysis from several months to less than one week. In the Alzheimer's disease domain, the CARE-AD framework, through collaboration among six specialized agents, achieved prediction of AD risk up to 10 years before formal diagnosis. Across these tasks, the evidence converges on a common conclusion: the value of multi-agent collaboration lies in reducing the cognitive burden and operational complexity of difficult tasks, rather than in improving the execution efficiency of simple ones[75]. Notably, different task types may require different collaboration architectures—diagnostic tasks depend more on single-round specialization, whereas research automation benefits more from multi-round iterative refinement[76]. This distinction suggests that future research should not pursue a one-size-fits-all architecture but instead design collaboration patterns tailored to the specific characteristics of each task[77].

Our findings indicate that the cost-effectiveness relationship between multi-agent systems and single models is not unidirectional. In the study by Yahya Shaikh et al.[26], the multi-agent system (five GPT-4 instances plus a facilitator) improved accuracy from 78.00% to 94.67%, and the cost per percentage point of accuracy decreased from $0.34 to $0.23, achieving better cost-effectiveness. In contrast, in the study by Xinti Sun et al.[25], the multi-agent system had a higher cost per

percentage point of accuracy ($0.28) than the three single models ($0.04–$0.11). This contrast suggests that architectural factors—such as debate rounds, coordination mechanisms, and model selection—may influence the cost-effectiveness ratio. Previous studies have shown that the balance between performance and cost in multi-agent systems is highly design-dependent: lightweight orchestration can reduce token consumption by up to 65 times compared to single agents under mixed-task clinical workloads[78]; in replicating health economic models, multi-agent systems achieved higher accuracy and transparency but required longer processing times[79]; and in literature screening, majority voting achieved an optimal trade-off between cost and performance[56]. Therefore, when the accuracy gain from multi-agent collaboration outweighs the additional computational overhead, its economic efficiency can surpass that of the single model; otherwise, it may not. Cost optimization in multi-agent systems is an achievable goal through architectural design, rather than a fixed constraint. Future research should explore optimization strategies such as lightweight deployment, task-selective activation, and model heterogeneity, and conduct systematic cost-utility analyses in real-world clinical settings to define the boundaries of economic feasibility.

The CHART guideline evaluation reveals systematic deficiencies in methodological transparency and reproducibility in this field. The lack of transparency in prompt engineering is the most prominent shortcoming—full prompts were reported in only 37.0% of studies, and prompt evolution processes in only 19.6%. In LLM research, prompts are the intervention; different wording, structure, and context can lead to substantially different outputs. Failing to report prompt details is equivalent to failing to report the dosing regimen in a randomized controlled trial[80]. Reporting rates for query date and location (6.5%) and independent session usage (15.2%) were similarly low. LLM outputs are inherently stochastic—the same prompt can yield different responses at different times or across different sessions. Not reporting the time and environmental conditions of queries makes study results essentially irreproducible[19].

For CHA research to generate high-level evidence for clinical decision-making, systematic standards need to be established in several areas: complete disclosure of prompt engineering and query environments, prospective study registration, and clear reporting of ethical approval and data privacy protection measures. Journal editors and peer reviewers should actively promote the adoption of relevant reporting guidelines and consider reporting completeness as a key dimension in manuscript evaluation.

Based on the findings of this study, we propose the following directions for future research.
First, methodological standardization: the field urgently needs systematic adoption of reporting guidelines such as CHART, REFINE, and TRIPOD-LLM, particularly for full disclosure of prompt engineering and querying environments, as well as preregistration of study protocols.
Second, upgrading study design: shift from retrospective benchmarking to prospective multicenter clinical trials that use real-world data to validate the safety, effectiveness, and cost-effectiveness of multi-agent systems.
Third, cost optimization strategies: explore technical approaches such as model lightweighting, selective task activation, and local deployment to lower the computational barriers of multi-agent systems.

Conclusion

This study provides a systematic overview of the landscape of multi-agent collaboration research in medicine through bibliometric analysis, systematic review, meta-analysis, CHART guideline evaluation, and cost analysis. The field has seen rapid growth in publications since 2024, with most articles published in JCR Q1 journals. However, the country distribution is highly concentrated in China and the United States (together accounting for 67.4%), with data from low-income countries absent; thus, the external validity of current findings requires validation across multiple regions. The included studies cover diverse areas such as medical question answering, disease diagnosis, research workflow automation, Alzheimer's disease, and clinical text generation. The meta-analysis confirmed that, in MedQA tasks, multi-agent systems showed higher performance LLMs ($p < 0.001$). However, most studies did not report question-level paired data; therefore, the conclusions should be considered evidence of efficacy under controlled conditions rather than evidence of effectiveness in real-world clinical settings. Cost analysis suggests that the cost-effectiveness relationship between multi-agent systems and single models is not unidirectional—appropriate cost increases can be traded for higher reliability—but systematic cost-utility analyses are currently lacking. The CHART guideline evaluation revealed systematic deficiencies in methodological transparency and reporting standards in this field, with reporting rates for key items such as prompt engineering, query environment, and preregistration generally below 40%. Moreover, existing quality assessment tools are unable to fully capture the specific sources of bias inherent in AI benchmark comparison studies. Future research should adopt standardized reporting guidelines, conduct prospective multicenter clinical trials, explore optimal collaboration architectures, establish standardized evaluation frameworks, and incorporate systematic cost-utility analyses, in order to move multi-agent systems from proof-of-concept toward true clinical decision support tools.

References:

[1] Bedi, S. *et al.*, Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA.* **333**, 319 (2025).

[2] Lin, C. ,Kuo, C., Roles and potential of Large language models in healthcare: A comprehensive review. *Biomed. J.* **48**, 100868 (2025).

[3] Z. L. Chen *et al.*, Applications and Prospects of Large Language Models in the Field of Healthcare. *医学与哲学* **46**, 32 (2025).

[4] Salehin, I. *et al.*, Systematic Literature Review of LLM-Large Language Model in Medical: Digital Health, Technology and Applications. *Eng. Rep.* **7**, e70365 (2025).

[5] Lee, Y., Wang, X. ,Yang, C. C., Automated Clinical Problem Detection from SOAP Notes using a Collaborative Multi-Agent LLM Architecture. *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (2025).

[6] Park, J. S. *et al.*, Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023).

[7] Chen, X. *et al.*, Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ Digit. Med.* **8**, 159 (2025).

[8] Jiang, Y., Chen, J., Liu, L. ,Clarkson, P. J., Exploring medical co-design in the AI era: A scoping review. *Design and Artificial Intelligence* **1**, 100040 (2025).

[9] Chen, X. *et al.*, RareAgents: Autonomous Multi-disciplinary Team for Rare Disease Diagnosis and Treatment. *ArXiv* **abs/2412.12475**, (2024).

[10] Miao, Y., Wen, J., Luo, Y. ,Li, J., MedARC: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *Int. J. Med. Inform.* **206**, 106136 (2026).

[11] Swanson, K. *et al.*, The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716 (2025).

[12] Tzanis, E. *et al.*, Agentic systems in radiology: Principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagn. Interv. Imaging* **107**, 7 (2026).

[13] Salehi, S. *et al.*, Systematic Review: Agentic AI in Neuroradiology: Technical Promise with Limited Clinical Evidence. *Journal of Imaging Informatics in Medicine* (2026).

[14] Kim, H., Kwon, T., Kim, S. ,Kim, S., Virtual lab of artificial intelligence agents accelerating nanobody design against SARS-CoV-2 variants. *Int. J. Biol. Sci.* **22**, 618 (2026).

[15] Radford, A., Narasimhan, K., Salimans, T. ,Sutskever, I., Improving Language Understanding by Generative Pre-Training, https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (2018).

[16] Chen, S. F. *et al.*, LLM-assisted systematic review of large language models in clinical medicine. *Nat. Med.* **32**, 1152 (2026).

[17] Landis, J. R. ,Koch, G. G., The measurement of observer agreement for categorical data. *Biometrics* **33**, 159 (1977).

[18] Kim, S. ,Yoon, H., in Questioning Our Questions: How Well Do Medical {QA} Benchmarks Evaluate Clinical Capabilities of Language Models? (Viena, Austria, 2025)pp. 274.

[19] Yao, G. *et al.*, Comparing the accuracy of large language models and prompt engineering in diagnosing realworld cases. *Int. J. Med. Inform.* **203**, 106026 (2025).

[20] Wang, J., Arora, K., Swoboda, D. M. ,Nazha, A., Tumor Board–Inspired Multiagent Artificial Intelligence System for Interpreting Oncology Guidelines. *JCO Clin. Cancer Inform.* **10**, e2500286 (2026).

[21] Deeks, J. J., Higgins, J. P., Altman, D. G. ,On, B. O. T. C., (2019), pp. 241.

[22] Haidich, A. B., Meta-analysis in medical research. *Hippokratia* **14**, 29 (2010).

[23] Stewart, L. A. ,Tierney, J. F., To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval. Health. Prof.* **25**, 76 (2002).

[24] Riley, R. D., Lambert, P. C. ,Abo-Zaid, G., in Meta-analysis of individual participant data: rationale, conduct, and reporting, (BMJ Publishing Group Ltd, 2010), Vol. **340},**

elocation-id = {c221.

[25] Sun, X. *et al.*, Model confrontation and collaboration: A debate intelligence framework for enhancing medical reasoning in large language models. *Cell Rep. Med.* **7**, 102547 (2026).

[26] Shaikh, Y. *et al.*, Collaborative intelligence in AI: Evaluating the performance of a council of AIs on the USMLE. *PLOS digital health* **4**, e787 (2025).

[27] Huo, B. *et al.*, Reporting guideline for Chatbot Health Advice studies: the CHART statement. *BMC Med.* **23**, 447 (2025).

[28] Wan, X. *et al.*, DeepGut: A collaborative multimodal large language model framework for digestive disease assisted diagnosis and treatment. *World J. Gastroenterol.* **31**, 109948 (2025).

[29] Mukhtar, A., Arzu, G. E., Toor, W. T. ,Ali, U., Collaborative multi-agent conversational artificial intelligence for clinical support in Parkinson disease. *Parkinsonism Relat. Disord.* 108292 (2026).

[30] Luo, Y. *et al.*, From intention to implementation: automating biomedical research via LLMs. *Science China Information Sciences* **68**, 170105 (2025).

[31] Cai, C. *et al.*, Analyzing the teaching and learning environments through student feedback at scale: a multi-agent LLMs framework. *Educ. Inf. Technol.* **30**, 21815 (2025).

[32] Wang, J., Swoboda, D. M. ,Nazha, A., Autonomous Analysis of Curated Patient Data Using a Large Language Model–Based Multiagent Framework. *JCO Clin. Cancer Inform.* e2500176 (2025).

[33] Sorka, M., Gorenshtein, A., Aran, D. ,Shelly, S., A multi-agent approach to neurological clinical reasoning. *PLOS digital health* **4**, e1106 (2025).

[34] Ekingen, E. ,ü dal, M., Multimodal AI-assisted Diagnosis and Management of Cardiac Emergencies: A Pilot Study. *Eurasian J. Emerg. Med.* (2026).

[35] Elmitwalli, S., Mehegan, J., Braznell, S. ,Gallagher, A., Development and validation of a multi-agent AI pipeline for automated credibility assessment of tobacco misinformation: a proof-of-concept study. *Front. Artif. Intell.* **Volume 8 - 2025**, (2025).

[36] De Maio, C. *et al.*, Privacy-Preserving Healthcare Data Interactions: A Multi-Agent Approach Using LLMs. *J. Commun. Softw. Syst.* **21**, 13 (2025).

[37] Xie, E. *et al.*, CASSIA: a multi-agent large language model for automated and interpretable cell annotation. *Nat. Commun.* **17**, 389 (2025).

[38] L. Han, G. Q. Liang ,J. Xu, Construction of Intelligent Agents for Ancient Chinese Medicine Knowledge Services Based on Large Language Models. *信息与管理研究* **11**, 31 (2026).

[39] Gorenshtein, A. *et al.*, Agent-guided AI-powered interpretation and reporting of nerve conduction studies and EMG (INSPIRE). *Clin. Neurophysiol.* **177**, 2110792 (2025).

[40] Gao, J. *et al.*, MoMA: a mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *NPJ Digit. Med.* **9**, 46 (2025).

[41] Liu, H. *et al.*, LLM-driven collaborative framework for knowledge-enhanced cancer pain assessment and management. *NPJ Digit. Med.* **9**, 180 (2026).

[42] Ucdal, M., Yurtsever, K. ,Ekingen, E., Multidisciplinary artificial intelligence systems versus single-model approaches for the diagnosis and management of ileus

and volvulus. *BMC Gastroenterol.* **26**, 124 (2026).

[43] Livieratos, A. *et al.*, MetaMind: A multi-agent transformer-driven framework for automated network meta-analyses. *PLoS One* **21**, e342895 (2026).

[44] Han, S. ,Choi, W., Development of a Large Language Model-based Multi-Agent Clinical Decision Support System for Korean Triage and Acuity Scale (KTAS)-Based Triage and Treatment Planning in Emergency Departments. *ArXiv* **abs/2408.07531**, (2024).

[45] Pan, C. A. L. W., Automated Literature Screening for Hepatocellular Carcinoma Treatment Through Integration of 3 Large Langua ge Models: Methodological Study. *JMIR Med. Inform.* **13**, e76252 (2025).

[46] Ho, A. Z. T. *et al.*, DILIConsult: A Multi-Agent Large Language Model Framework for Evaluating Drug-Induced Liver Injury in ICU Settings. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **46**, e70131 (2026).

[47] Li, R. *et al.*, CARE-AD: a multi-agent large language model framework for Alzheimer’s disease prediction using longitudinal clinical notes. *NPJ Digit. Med.* **8**, 541 (2025).

[48] Liu, S. *et al.*, Optimizing Order Sets With a Large Language Model–Powered Multiagent System. *JAMA Netw. Open* **8**, e2533277 (2025).

[49] P. Wang *et al.*, Research and Implementation of a Rare Disease Intelligent Discovery System Driven by Large Language Models. *中国数字医学* **21**, 30 (2026).

[50] Chen, Y., Wen, B. ,Zulkernine, F., A Multiagent Summarization and Auto-Evaluation Framework for Medical Text: Development and Evaluation Study. *JMIR AI* **4**, e75932 (2025).

[51] Peng, Q. *et al.*, Integration of Multi-Source Medical Data for Medical Diagnosis Question Answering. *IEEE Trans. Med. Imaging* **44**, 1373 (2025).

[52] dü, A. A. U., Arslano lu, K. ,Karak se, M., An Adaptive Multi-Agent LLM-Based Clinical Decision Support System Integrating Biomedical RAG and Web Intelligence. *IEEE Access* **13**, 167390 (2025).

[53] Tian, J. *et al.*, An autonomous agentic workflow for clinical detection of cognitive concerns using large language models. *NPJ Digit. Med.* **9**, 51 (2026).

[54] Hasjim, B. J. *et al.*, A multiagent large language model-based system to simulate the liver transplant selection committee: a retrospective cohort study. *The Lancet Digital Health* **8**, 100966 (2026).

[55] Jiang, Z. ,Feng, S., UsmleGPT: An AI application for developing MCQs via multi-agent system. *Software Impacts* **23**, 100742 (2025).

[56] Akinseloyin, O., Jiang, X. ,Palade, V., Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. *Biology Methods and Protocols* **11**, bpag6 (2026).

[57] Xu, Q. *et al.*, Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience* **28**, 113984 (2025).

[58] Liu, B. *et al.*, Magic: AN LLM-based multi-agent activated graph-reasoning intelligent collaboration model for liver disease diagnosis. *Inf. Fusion* **126**, 103557 (2026).

[59] Hedaya, R. J., Iterative Dual-AI Consultation for Error Detection in Clinical Medicine: A Case Study Demonstrating Convergent Validity Through Cross-Validation of Large Language Models. *Altern. Ther. Health Med.* **32**, (2026).

[60] Khan, M. A. *et al.*, Collaborative large language models for automated data extraction in living systematic reviews. *Journal of the American Medical Informatics Association : JAMIA* **32**, 638 (2025).

[61] Gui, X. *et al.*, Enhancing hepatopathy clinical trial efficiency: a secure, large language model-powered pre-screening pipeline. *BioData Min.* **18**, 42 (2025).

[62] Gallingani, C. *et al.*, in Agentic Generative Artificial Intelligence System for Classification of Pathology-Confirmed Primary Progressive Aphasia Variants}, elocation-id = {2025.10.28.25338977, (Cold Spring Harbor Laboratory Press, 2025).

[63] Altermatt, F. R. *et al.*, Performance of single-agent and multi-agent language models in Spanish language medical competency exams. *BMC Med. Educ.* **25**, 666 (2025).

[64] Li, J. *et al.*, Experience-guided multi-agent interpretable framework for radiology report summarization. *Comput. Methods. Programs. Biomed.* **273**, 109078 (2026).

[65] Zhang, M. *et al.*, A large language model-based self-learning and critical agent framework for multimodal Alzheimer’s disease diagnosis. *Neuropsychology* (2026).

[66] Y. Liu *et al.*, Integration of Knowledge Graph and Multi-Character LLM in Alzheimer's Question Answering Research. *计算机应用与软件* 1.

[67] Liu, W. *et al.*, Detecting children with autism spectrum disorder based on script-centric behavior understanding with emotional enhancement. *Neurocomputing* **670**, 132434 (2026).

[68] H. W. Zhou, X. K. Liang ,Y. Q. Xie, OncMDT: Dynamic Composition and Reflective Coordination of Multi-Agent Framework for Tumor Multidisciplinary Consultation. *集成技术* 1.

[69] Klingelhöfer, D. *et al.*, Research on artificial intelligence, machine and deep learning in medicine: global characteristics, readiness, and equity. *Global. Health* **21**, 36 (2025).

[70] Guo, S. *et al.*, Surge in large language models exacerbates global regional healthcare inequalities. *J. Transl. Med.* **23**, 706 (2025).

[71] Chen, Z. A. L. W., Characterizing patients who benefit from mature medical AI models in real-world clinical applications. *PLOS Digital Health* **5**, 1 (2026).

[72] V A Squez-Venegas, C. *et al.*, in Who is leading medical AI? A systematic review and scientometric analysis of chest x-ray research}, elocation-id = {2026.04.02.26349884, (Cold Spring Harbor Laboratory Press, 2026).

[73] Gong, E. J., Bang, C. S., Lee, J. J. ,Baik, G. H., Knowledge-Practice Performance Gap in Clinical Large Language Models: Systematic Review of 39 Benchmarks. *J. Med. Internet Res.* **27**, e84120 (2025).

[74] Bhayana, R. *et al.*, From Bench to Bedside With Large Language Models: AJR Expert Panel Narrative Review. *AJR Am. J. Roentgenol.* **223**, e2430928 (2024).

[75] Xu, X. ,Sankar, R., Large Language Model Agents for Biomedicine: A Comprehensive Review of Methods, Evaluations, Challenges, and Future Directions. *Information* **16**, 894 (2025).

[76] Spieser, J. *et al.*, A Review of Multi-Agent AI Systems for Biological and Clinical Data Analysis. *Methods and Protocols* **9**, 33 (2026).

[77] Lin, Y. *et al.*, A Survey of LLM-based Multi-agent Systems in Medicine. *TechRxiv* **2025**, (2025).

[78] Klang, E. *et al.*, Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *npj Health Systems* **3**, 23 (2026).

[79] Samur, S. *et al.*, MSR188 Single-Agent vs. Multi-Agent RAG for Health Economic Model Replication: A Comparative Evaluation. *Value Health* **28**, S532 (2025).

[80] Collaborative, T. C., Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *BJS* **112**, znaf142 (2025).

Supplementary Files

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

- CostAnalysis.xlsx
- Supplementarymaterials.docx