Article:
Akinseloyin, O., Jiang, X. and Palade, V. (2026) Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. Biology Methods and Protocols, 11 (1). bpag006. ISSN: 2396-8923

https://doi.org/10.1093/biomethods/bpag006

Biology Methods and Protocols, 2026, **11**(1), bpag006
https://doi.org/10.1093/biomethods/bpag006
Published: 4 February 2026
Methods Article

BIOLOGY
Methods & Protocols

OXFORD

Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews

Opeoluwa Akinseloyin[1], Xiaorui Jiang[2,*] and Vasile Palade[1]

[1]Centre for Computational Science and Mathematical Modelling, Coventry University, Puma Way, Coventry, CV1 2TT, United Kingdom
[2]School of Information, Journalism and Communications, The University of Sheffield, The Wave, 2 Whitham Rd, Sheffield, S10 2AH, United Kingdom

*Corresponding author. School of Information, Journalism and Communications, The University of Sheffield, The Wave, 2 Whitham Rd, Sheffield, S10 2AH, United Kingdom.
E-mail: xiaorui.jiang@sheffield.ac.uk

Abstract

Systematic reviews (SRs) are essential for evidence-based practice but remain labor-intensive, especially during abstract screening. This study evaluates whether multiple large language models (multi-LLMs) collaboration can improve the efficiency and reduce costs for abstract screening. Abstract screening was framed as a question-answering (QA) task using cost-effective LLMs. Three multi-LLM collaboration strategies were evaluated, including majority voting by averaging opinions of peers, multi-agent debate for answer refinement, and LLM-based adjudication against answers of individual QA baselines. These strategies were evaluated on 28 SRs of the CLEF eHealth 2019 technology-assisted review benchmark using standard performance metrics such as mean average precision (MAP) and work saved over sampling at 95% recall (work saved over sampling WSS@95%). Multi-LLM collaboration significantly outperformed QA baselines. Majority voting was overall the best strategy, achieving the highest MAP 0.462 and 0.341 on subsets of SRs about clinical intervention and diagnostic technology assessment, respectively, with WSS@95% 0.606 and 0.680, enabling in theory up to 68% workload reduction at 95% recall of all relevant studies. Multi-agent debate improved weaker models most. Our own adjudicator-as-a-ranker method was the second strongest approach, surpassing adjudicator-as-a-judge, but at a significantly higher cost than majority voting and debating. Multi-LLM collaboration substantially improves abstract screening efficiency, and the success lies in model diversity. Making the best use of diversity, majority voting stands out in terms of both excellent performance and low cost compared to adjudication. Despite context-dependent gains and diminishing model diversity, multi-agent debate is still a cost-effective strategy and a potential direction of further research.

Keywords systematic review, abstract screening, large language model, ensemble, multiagent system

Introduction

Systematic reviews serve as cornerstones for evidence-based practice, particularly in medicine and healthcare, by providing rigorous summaries of existing knowledge [1]. However, conducting SRs is notoriously labor-intensive, often requiring researchers to spend over several months [2, 3]. A primary bottleneck is the title and abstract screening phase, where the sheer volume of retrieved studies—sometimes as high as tens of thousands [4, 5]—presents formidable challenges, amplified by the standard practice of involving multiple human annotators [6].

There have been two decades of efforts on using AI to automate or semi-automate the screening task since Cohen et al.'s seminal works [7, 8]. Early automation efforts leveraged machine learning techniques [2, 9], later progressing to deep learning models [10, 11]. These approaches faced significant challenges, such as the requirement for extensive labeled data, extreme class imbalance and the need for model retraining for each new review [2, 12–15]. Active learning sought to alleviate these issues through iterative human feedback [16–19], but faced challenges in achieving a high degree of automation due to the zero-shot nature of this problem [20, 21].

The advent of large language models (LLMs) marked a paradigm shift, offering unprecedented zero-shot learning capabilities for tasks like abstract screening [22–27]. Recent studies have demonstrated LLMs' potential across various SR stages, from search query generation [28] to literature screening tasks [29–32] and even the whole SR pipeline [33]. As shown in Michiel et al. [19] and Akinseloyin et al. [34], LLMs can also make the active learning process for abstract screening more efficient. However, single LLMs inevitably suffer inherent model biases [35, 36] and weak alignment with nuanced

Received: 24 September 2025. **Revised:** 18 January 2026. **Accepted:** 19 January 2026

human judgment [37], making it more less likely for individual LLMs to meet the sensitivity/recall requirement and reach a good balance between sensitivity and specificity [38]. These limitations have catalyzed interest in enhancing reliability through multiple LLM (multi-LLM) collaboration, aka LLM-based multiagent systems (MASs) [39–41]. For example, ensemble approaches, a primitive form of multimodel collaboration, have demonstrated superior performance for abstract screening by combining the decisions from a range of LLMs [38, 42, 43] to achieve higher sensitivity/recall and better balance between sensitivity and specificity/precision.

LLM-based MAS (here agent is an LLM) has emerged as a powerful paradigm across various domains, with extensive research demonstrating their effectiveness [39, 44, 45]. Key collaboration strategies include debating mechanisms [46, 47], where agents engage in structured argumentation [45, 46, 48], and adjudication approaches leveraging, for example, the LLM-as-a-Judge frameworks [49, 50]. These approaches aim to mitigate individual model weaknesses through collective reasoning among peers and have been demonstrated effective for complex reasoning tasks in medical domains [51–53]. Yet, a comprehensive investigation of multi-LLM collaboration hasn't been presented in the context of abstract screening. This paper embarks on the first comprehensive investigation into multi-LLM collaboration in automated abstract screening. Our contributions are threefold: (i) We investigate multi-LLM collaborative strategies, including ensembling, debating, and adjudication, toward SR automation. (ii) We comparatively evaluate different strategies to identify the most robust approaches. (iii) We empirically analyze the core success factors that enable multi-LLM collaboration to mitigate errors of individual LLMs.

Methodology

LLM-based question answering for screening

Consistent with Akinseloyin et al. [30], we formulate the abstract screening task using a question-answering (QA) framework. Suppose each SR constitutes an unannotated dataset of (the titles and abstracts of) candidate studies (i.e. documents) $D = \{d_1, d_2, \ldots, d_N\}$, where N is the total number of documents and di is the i-th document. Abstract screening is the process of assigning a label to indicate that a document should be "included" into or "excluded" from the remaining steps of an SR, for which screening prioritization ranks the documents in descending order of their likelihood of being included.

Each SR has a paragraph of selection criteria questions, $Q = \{q_1, \ \cdots, \ q_K\}$, that every included study must satisfy. Given a document d_i $(i = 1, \ldots, \ N)$, each inclusion criteria question q_k $(k = 1, \ldots, K)$ will be answered by an LLM-based QA model $\mathbf{M}^{\mathrm{qa}}$, with the answer $a_{i,k}^{\mathbf{M}^{\mathrm{qa}}}$ being either "Positive" (meaning meeting the criterion), "Negative" (meaning not) or "Neutral" (meaning unsure or not answerable) plus a reasoning text. To ease discussion, the QA process is formalized as follows, with the prompt and a corresponding example presented in Supplementary Appendices A and B of Supplementary File 1 respectively:

$$\mathcal{O}_{i,k}^{\mathbf{M}^{\mathrm{qa}}} = \mathbf{M}^{\mathrm{qa}}(\mathcal{I}_{i,k}^{\mathrm{qa}}),$$

where the input $\mathcal{I}_{i,k}^{\mathrm{qa}} = \langle q_k, d_i \rangle$ contains the k-th inclusion criteria question on the i-th document, and the output $\mathcal{O}_{i,k}^{\mathbf{M}^{\mathrm{qa}}} = \left\langle a_{i,k}^{\mathbf{M}^{\mathrm{qa}}}, \ r_{i,k}^{\mathbf{M}^{\mathrm{qa}}} \right\rangle$ contains the answer $a_{i,k}^{\mathbf{M}^{\mathrm{qa}}}$ and its reasoning $r_{i,k}^{\mathbf{M}^{\mathrm{qa}}}$ for the k-th question on the i-th document.

Given a QA model $\mathbf{M}^{\mathrm{qa}}$, we use the same method in Akinseloyin et al. [30] to score d_i with respect to each question q_k, by assigning the probability that the corresponding answer text (i.e. the concatenation of $a_{i,k}^{\mathbf{M}^{\mathrm{qa}}}$ and $r_{i,k}^{\mathbf{M}^{\mathrm{qa}}}$ has a positive sentiment, according to a pretrained BART model) [54]:

$$\mathrm{score}\left(d_i, q_k; \mathbf{M}^{\mathrm{qa}}\right) = \mathrm{Prob}_{\mathrm{BART}}\left(\mathrm{positive}|a_{i,k}^{\mathbf{M}^{\mathrm{qa}}}, \ r_{i,k}^{\mathbf{M}^{\mathrm{qa}}}\right). \qquad (1)$$

The document score is the sum of its scores with respect to all inclusion criteria:

$$\mathrm{score}\left(d_i, \mathcal{Q}; \mathbf{M}^{\mathrm{qa}}\right) = \sum\nolimits_{k=1}^{K} \mathrm{score}\left(d_i, q_k; \mathbf{M}^{\mathrm{qa}}\right). \qquad (2)$$

Multi-LLM collaboration

Inspired by recent work in LLM-based MAS [39, 46, 55], we investigate three distinct strategies for combining the outputs and reasoning from multiple LLMs (aka agents). The prompts for all strategies are presented in Supplementary Appendix A in Supplementary File 1, while Supplementary Appendix B presents an illustrative example for each of the three strategies.

Majority voting

Assuming the "wisdom of the crowd" supersedes individuals, the first simple but extremely effective strategy is majority voting, more precisely soft voting (Soft-Vote). Given a collection of primary (QA) models $\mathcal{M} = \{\mathbf{M}_l\}|_{l=1}^{L}$, for each question q_k and document di, the soft voting score is defined as the average of the scores for each primary model:

$$\mathrm{score}_{\mathrm{vote}}\left(d_i, q_k; \mathcal{M}\right) = \frac{1}{L}\sum\nolimits_{l=1}^{L} \mathrm{score}_{\mathrm{vote}}\left(d_i, q_k; \mathbf{M}_l\right). \qquad (3)$$

Then the final score for each document is the sum of the soft voting scores with respect to all inclusion criteria:

$$\mathrm{score}_{\mathrm{vote}}(d_i, \mathcal{Q}; \mathcal{M}) = \sum\nolimits_{k=1}^{K} \mathrm{score}_{\mathrm{vote}}\left(d_i, q_k; \mathcal{M}\right). \qquad (4)$$

Multiagent debate

This strategy introduces a step of cross-agent communication, reflection, and refinement [46, 56]. After an initial round of independent QA, each agent is presented with the answers and reasoning of other agents for the same question on an abstract, and is prompted to reconsider its initial answer and reasoning in light of the perspectives of its peers. If an answer is changed, the updated answer and reasoning are recorded.

Formally, the debating process is defined as follows:

$$\mathcal{O}_{i,k}^{\mathbf{M}_l^{\mathrm{deb}}} = \mathbf{M}_i^{\mathrm{deb}}(\mathcal{I}_{i,k}^{\mathbf{M}_l^{\mathrm{deb}}}),$$

where $\mathbf{M}_i^{\mathrm{deb}} \in \mathcal{M}$ $(\mathrm{l} = 1, \ldots, \mathrm{L})$ is the debating model, the input $\mathcal{I}_{i,k}^{\mathbf{M}_i^{\mathrm{deb}}}$ is the union of the output of itself and the outputs of all other QA models, plus the original QA input, that is,

$$\mathcal{I}_{i,k}^{\mathbf{M}_l^{\text{deb}}} = \left\langle \mathcal{I}_{i,k}^{\text{qa}}, \mathcal{O}_{i,k}^{\mathbf{M}_l^{\text{qa}}}, \mathcal{O}_{i,k}^{\mathbf{M}_{l'}^{\text{qa}}} \Big|_{l'=1, l'\neq l}^{L} \right\rangle,$$

and the output of the debating model is as follows

$$\mathcal{O}_{i,k}^{\mathbf{M}_l^{\text{deb}}} = \left\langle a_{i,k}^{\mathbf{M}_l^{\text{deb}}}, r_{i,k}^{\mathbf{M}_l^{\text{deb}}}, v_{i,k}^{\mathbf{M}_l^{\text{deb}}}, c_{i,k}^{\mathbf{M}_l^{\text{deb}}}, e_{i,k}^{\mathbf{M}_l^{\text{deb}}} \right\rangle,$$

with $a_{i,k}^{\mathbf{M}_l^{\text{deb}}}$, $r_{i,k}^{\mathbf{M}_l^{\text{deb}}}$, and $v_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ being the new ***a***nswer, the ***r***easoning, and the confidence ***v***alue of the debating agent on the new answer, $c_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ indicating whether the debating agent ***c***hanges the original answer ("Yes" or "No"), and $e_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ holding the ***e***xplanation for why the original answer is changed or kept unchanged. Note that $\mathbf{M}_l^{\text{deb}}$ and $\mathbf{M}_l^{\text{qa}}$ refer to the same model $\mathbf{M}_l \in \mathcal{M}$ used for different purposes, taking different inputs and generating different outputs. Note that the new reasoning $r_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ is the updated rationale for the current answer, while the explanation $e_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ explicitly states why the answer was changed or kept unchanged after considering peers' input. See Supplementary Appendices C and D in Supplementary File 1 for an example and further explanations.

The final scoring and ranking are based on the answers after debating. Formally, given a collection of LLMs, amongst which $\mathbf{M}_l^{\text{deb}}$ is the debating model, the debating score for each document is defined as follows:

$$\text{score}_{\text{debate}}\left(d_i, \mathcal{Q}; \mathbf{M}_l^{\text{deb}}\right) = \sum_{k=1}^{K} \text{score}_{\text{debate}}\left(d_i, q_k; \mathbf{M}_l^{\text{deb}}\right). \quad (5)$$

where

$$\text{score}_{\text{debate}}\left(d_i, q_k; \mathbf{M}_l^{\text{deb}}\right) = \text{Prob}_{\text{BART}}\left(\text{Positive}|a_{i,k}^{\mathbf{M}_l^{\text{deb}}}\right). \quad (6)$$

Note that here the new reasoning $r_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ is not used for scoring because the debating model may change the answer and the explanation for such change may contain negative information against the old answer rather than negative viewpoints against the question, in which case the linguistic cues for answer justification will mislead BART in score assignment. Instead, scoring is based solely on the final answer $a_{i,k}^{\mathbf{M}_l^{\text{deb}}}$. In cases where documents receive identical scores, the debating agent's confidence $v_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ is used to break ties by producing a weighted score, ensuring that answers supported with higher certainty are prioritized in the final ranking. Also note that our multiagent debate (MAD) approach, together with LLM-based adjudication to be introduced below, can be seen as variants of the idea of exchange-of-thought [48].

LLM-based adjudication

The third strategy uses a separate and more powerful LLM as the adjudicator to synthesize different opinions and make the final verdict [57]. For each question, the adjudicator receives the initial answers of all primary models, analyzes their reasoning texts, and determines which answer is the most accurate or well-justified. Formally, given a collection of LLMs $\mathcal{M}$ as QA models, for each inclusion criteria question q_k on a document d_i, the judging process of the adjudicator LLM $\mathbf{M}^{\text{adj}} \notin \mathcal{M}$ is defined as follows:

$$\mathcal{O}_{i,k}^{\mathbf{M}^{\text{adj}}} = \mathbf{M}^{\text{adj}}\left(\mathcal{I}_{i,k}^{\mathbf{M}^{\text{adj}}}\right),$$

where the input $\mathcal{I}_{i,k}^{\mathbf{M}^{\text{adj}}}$ is exactly the same as $\mathcal{I}_{i,k}^{\mathbf{M}_l^{\text{deb}}}$ and the output is defined as follows,

$$\mathcal{O}_{i,k}^{\mathbf{M}^{\text{adj}}} = \left\langle a_{i,k}^{\mathbf{M}^{\text{adj}}}, r_{i,k}^{\mathbf{M}^{\text{adj}}}, v_{i,k}^{\mathbf{M}^{\text{adj}}}, g_{i,k}^{\mathbf{M}_l^{\text{qa}}} \Big|_{l=1}^{L}, b_{i,k}^{\mathbf{M}^{\text{qa}}}, w_{i,k}^{\mathbf{M}^{\text{qa}}} \right\rangle,$$

with $a_{i,k}^{\mathbf{M}^{\text{adj}}}$, $r_{i,k}^{\mathbf{M}^{\text{adj}}}$, and $v_{i,k}^{\mathbf{M}^{\text{adj}}}$ being the adjudicator's new ***a***nswer, its ***r***easoning and confidence ***v***alue, each $g_{i,k}^{\mathbf{M}_l^{\text{qa}}}$ being the grading of a QA model (a rating value between 0 and 1), and $b_{i,k}^{\mathbf{M}^{\text{qa}}}$ and $w_{i,k}^{\mathbf{M}^{\text{qa}}}$ being the best and worst models selected by the adjudicator, respectively.

Two variants are proposed.

Adjudicator as a judge

The first variant is to use this LLM adjudicator as a "Judge" [49], or called "meta-agent" in other literature [47]. Similar to other methods, the score of a document di with respect to each criteria question q_k by an adjudicator as a judge is the probability that its answer text (i.e. $a_{i,k}^{\mathbf{M}^{\text{adj}}} + r_{i,k}^{\mathbf{M}^{\text{adj}}}$) has a positive sentiment based on BART:

$$\text{score}_{\text{judge}}\left(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}\right) = \text{Prob}_{\text{BART}}\left(\text{Positive}|a_{i,k}^{\mathbf{M}^{\text{adj}}}\right), \quad (7)$$

and the final score of each document is:

$$\text{score}_{\text{judge}}\left(d_i, \mathcal{Q}; \mathbf{M}^{\text{adj}}, \mathcal{M}\right) = \sum_{k=1}^{K} \text{score}_{\text{judge}}\left(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}\right). \quad (8)$$

Adjudicator as a ranker

Alternatively, a separate LLM (called adjudicator) is asked to rate the quality of individual answers and generate a grade for each primary model, denoted by $g_{i,k}^{\mathbf{M}_l^{\text{qa}}}$ $(l = 1, \cdots, L)$, which are then used to calculate a weighted average of the primary models' scores, i.e. the score of a document d_i with respect to criteria question q_k by the adjudicator:

$$\text{score}_{\text{rank}}\left(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}\right) = \frac{1}{L}\sum_{l=1}^{L} g_{i,k}^{\mathbf{M}_l^{\text{qa}}} \cdot \text{score}\left(d_i, q_k; \mathbf{M}_l^{\text{qa}}\right). \quad (9)$$

Then, the final score for each document according to the adjudicator as a ranker is:

$$\text{score}_{\text{rank}}\left(d_i, \mathcal{Q}; \mathbf{M}^{\text{adj}}, \mathcal{M}\right) = \sum_{k=1}^{K} \text{score}_{\text{rank}}\left(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}\right). \quad (10)$$

Re-ranking

Screening prioritization performance can be further improved through re-ranking as in Akinseloyin et al. [30]. The rationale is an included study should meet all selection criteria or most of them (when there are certain criteria unanswerable), so we can expect a high semantic relevance between the requirements of an SR's inclusion criteria and the information in an included study.

Macro-level re-ranking (rr-mac)

Relevance is measured by the cosine similarity between the text embeddings of each candidate study and the selection criteria paragraph, denoted by rel(d_i, $\mathcal{Q}$). A macro-level re-ranking score is calculated as follows:

$$\text{score}_{\text{rr-mac}}^{*}(d_i, \mathcal{Q}) = (1-\alpha)\cdot \text{score}(d_i, \mathcal{Q}) + \alpha \cdot \text{rel}(d_i, \mathcal{Q}), \quad (11)$$

where $\alpha \in (0, 1)$, and score$(d_i, \mathcal{Q})$ is the score of d_i that is calculated according to either Equation (4), Equation (5), Equation (8), or Equation (10).

Micro-level re-ranking (rr-mic)

Cosine similarity is calculated between each included study and each inclusion criterion $q_k \in \mathcal{Q}$, denoted by $\mathrm{rel}(d_i, q_k)$. Micro-level re-ranking first calculates a new score for each document with respect to each question as follows:

$$\mathrm{score}^{*}_{\mathrm{rr-mic}}(d_i, q_k) = (1-\beta) \cdot \mathrm{score}(d_i, q_k) + \beta \cdot \mathrm{rel}(d_i,\ q_k), \qquad (12)$$

where $\beta \in (0,\ 1)$ and $\mathrm{score}(d_i, q_k)$ is calculated according to either Equation (3), Equation (6), Equation (7) or Equation (9). Then, a new score for each document is calculated by summing up the scores with respect to all criteria questions:

$$\mathrm{score}^{*}_{\mathrm{rr-mic}}(d_i, \mathcal{Q}) = \sum_{k=1}^{K} \mathrm{score}^{*}_{\mathrm{rr-mic}}(d_i, q_k). \qquad (13)$$

Experimental setup

Dataset and evaluation metrics

Evaluation is done on CLEF eHealth 2019 Task 2: Technology-Assisted Reviews in Empirical Medicine (TAR2019)—a famous standard benchmark for evaluating abstract screening methods, including 20 Cochrane reviews about clinical intervention (Intervention; in total 39 792 documents) and 8 reviews about diagnostic technology assessment (DTA; 26 830). For all experiments in this study, each document's title was concatenated with its abstract before being passed to the LLM. This ensured that inclusion cues present in titles such as study design, population characteristics, or intervention types were captured alongside the more detailed information typically found in abstracts. For the small subset of records in TAR2019 that contained only titles without abstracts, we processed these using the title text alone. Across all reviews, 66 622 documents were screened, with inclusion rates ranging from 0.2% to 36.1% (mean: 5.8%). Detailed per-review statistics, including exact document counts, inclusion rates, selection criteria paragraphs, and generated question sets for each review, are provided in Supplementary Appendix D of Supplementary File 1.

We employ standard TAR evaluation metrics, including the rank of the *L*ast *Rel*evant document (L_{Rel}), mean average precision (MAP), *R*ecall at top *k*% of documents screened (*R*@*k*%, for $k = 5, 10, 20, \ldots,$ 50), and *w*ork *s*aved over *s*ampling (WSS) at the recall level of *R*% (WSS@*R*%, for *R* = 95, 100) [7]. Metrics are calculated per SR and then averaged across all the SRs in each of the two categories.

LLMs and baselines

LLM selection is based on two factors: balance between capability and computational cost, and popularity of model family. In the experiments, the "lightweight" versions of the GPT, Gemini and Claude families are chosen:

- GPT-4o Mini (gpt-4o-mini-2024-07-18),
- Claude 3 Haiku (claude-3-haiku-20240307), and
- Gemini 1.5 Flash (gemini-1.5-flash-preview-0514).

The LLM adjudicator, acting as both the "Judge" agent and the "Ranker" agent, is:

- Gemini 1.5 Pro (gemini-1.5-pro-preview-0514).

To ensure reproducibility, the temperature parameters for all models are set to zero. The models for evaluation include:

- Three primary QA models according to Equation (2), named by `GPT`, `Haiku`, and `Gemini`;
- The soft voter according to Equation (4), named by `Soft-Vote`;
- The three MAD models according to Equation (5), named by `GPT-MAD`, `Haiku-MAD`, and `Gemini-MAD`;
- The adjudicator-as-a-judge (`Adj-Judge`) and adjudicator-as-a-ranker (`Adj-Rank`) methods according to Equation (8) and (10) respectively.

We also compare the re-ranking variants of the aforementioned approaches that integrate both macro- and micro-level re-ranking, according to Equations (11) and (13). The GPT embedding model "text-embedding-ada-002" is used to generate text embeddings for relevance calculation.

To ensure reproducibility, all prompts used in this study are provided in Supplementary Appendix A of Supplementary File 1. All experiments used a temperature parameter of 0.1 to maximize reproducibility and deterministic outputs except Gemini 1.5 Flash. Gemini 1.5 Flash does not allow a temperature of 0.0 on certain prompts/abstracts, so 0.1 was selected as the lowest consistent value across all models in our experiments. Implementation code, model parameters, and evaluation scripts are available at the following link under the MIT License: https://github.com/Ope-Akinseloyin/Multi_LLM-Citation-Screening.git.

Results and discussion

Main results: screening performance enhancement

Table 1 presents the performance comparisons, which are summarised in the following subsections.

Majority voting

On both DTA and Intervention, `Soft-Vote` substantially outperformed all individual QA baselines with statistically significant improvements in MAP (paired *t*-test: $P < 0.001$ for all pairwise comparisons; Wilcoxon signed-rank test: $P < 0.001$) and WSS@95% ($P < 0.01$ for all comparisons), obtaining the highest MAP, recall and WSS values among all competitors. Mean R@10% reached 64.67% (±28.37) and 66.71% (± 25.78) on DTA and Intervention respectively, which demonstrated this LLM ensemble's strong ranking capability for abstract screening. In term of theoretical workload savings, it achieved 0.680 (±0.228) WSS@95% and 0.667 (±0.266) WSS@100% on DTA, and 0.606% (± 0.219) WSS@95% and 0.527 (±0.270) WSS100%. Considering its low cost, the simple approach `Soft-Vote` is both strong and cost-effective (refer to Supplementary Appendix E in Supplementary File 1 for the detailed breakdown for human cost estimation). The simple aggregation effectively mitigates individual model weaknesses and biases. The theoretical average workload reductions can be as high as approximately 68.0% and 60.6% on the 8 DTA and 20 Intervention SRs, respectively when the recall threshold of relevant studies is 95%, satisfying a widely-adopted critical requirement for avoiding causing significant damage to the reliability of the conclusions of an SR.

Table 1 Performances of multi-LLM collaboration.

Setting	Models	MAP	R@5%	R@10%	R@20%	R@50%	WSS@95%	WSS@100%
DTA	GPT	0.271 (±0.170)	41.74% (±28.51)	56.41% (±29.37)	68.45% (±25.88)	89.65% (±14.26)	0.428 (±0.247)	0.362 (±0.294)
	Gemini	0.266 (±0.189)	47.86% (±31.09)	60.76% (±28.15)	75.46% (±26.17)	92.24% (±0.116)	0.529 (±0.297)	0.440 (±0.354)
	Haiku	0.182 (±0.145)	33.21% (±28.96)	46.99% (±24.90)	70.99% (±22.91)	90.15% (±12.73)	0.407 (±0.178)	0.288 (±0.224)
	Soft-Vote	0.341 (±0.166)	46.92% (±29.68)	**64.67% (±28.37)**	**79.31% (±25.67)**	**94.84% (±8.83)**	**0.680 (±0.228)**	**0.667 (±0.266)**
	GPT-MAD	0.271 (±0.165)	40.47% (±27.61)	59.78% (±29.63)	73.21% (±27.57)	90.22% (±13.64)	0.534 (±0.306)	0.378 (±0.367)
	Gemini-MAD	0.276 (±0.184)	44.99% (±28.14)	61.80% (±27.90)	77.12% (±27.80)	92.08% (±12.95)	0.573 (±0.272)	0.445 (±0.371)
	Haiku-MAD	0.286 (±0.177)	47.91% (±33.84)	61.27% (±30.40)	77.12% (±25.79)	91.14% (±13.81)	0.540 (±0.300)	0.450 (±0.363)
	MAD-Soft-Vote	0.328 (±0.179)	49.11% (±33.63)	64.66% (±28.67)	77.19% (±28.13)	91.65% (±12.19)	0.579 (±0.282)	0.456 (±0.359)
	Adj-Judge	0.284 (±0.169)	48.20% (±33.39)	64.31% (±30.51)	76.25% (±28.10)	91.26% (±12.46)	0.550 (±0.284)	0.460 (±0.343)
	Adj-Rank	0.345 (±0.176)	**49.79% (±31.50)**	64.22% (±30.66)	79.06% (±27.25)	93.57% (±10.28)	0.593 (±0.279)	0.500 (±0.344)
	Adj-Soft-Vote	**0.352 (±0.178)**[1]	49.43% (±31.67)	64.63% (±29.93)	78.59% (±27.94)	93.57% (±10.28)	0.594 (±0.279)	0.515 (±0.333)
Intervention	GPT	0.395 (±0.265)	46.01% (±26.61)	62.32% (±24.82)	77.63% (±19.13)	90.87% (±13.70)	0.464 (±0.311)	0.392 (±0.329)
	Gemini	0.389 (±0.248)	48.57% (±28.86)	62.49% (±24.86)	76.24% (±17.82)	93.57% (±9.36)	0.481 (±0.247)	0.413 (±0.296)
	Haiku	0.290 (±0.241)	35.49% (±22.56)	52.64% (±24.36)	73.91% (±16.44)	93.74% (±6.38)	0.430 (±0.237)	0.344 (±0.287)
	Soft-Vote	***0.462 (±0.262)***	53.15% (±29.31)	66.71% (±25.78)	**83.41% (±15.74)**	**96.82% (±5.63)**	**0.606 (±0.219)**	0.527 (±0.270)
	GPT-MAD	0.376 (±0.267)	48.50% (±25.47)	62.06% (±24.59)	75.57% (±21.86)	92.30% (±9.51)	0.449 (±0.302)	0.389 (±0.319)
	Gemini-MAD	0.402 (±0.296)	49.87% (±27.00)	63.37% (±22.40)	80.21% (±19.23)	93.23% (±9.59)	0.509 (±0.278)	0.439 (±0.308)
	Haiku-MAD	0.419 (±0.263)	**54.43% (±30.03)**	68.01% (±24.73)	81.30% (±20.58)	96.02% (±7.18)	0.599 (±0.168)	**0.536 (±0.208)**
	MAD-Soft-Vote	0.456 (±0.272)	53.44% (±28.38)	**68.49% (±25.97)**	82.05% (±14.68)	96.28% (±7.39)	0.589 (±0.252)	0.527 (±0.312)
	Adj-Judge	0.427 (±0.269)	53.17% (±28.72)	67.29% (±24.10)	80.19% (±18.17)	93.19% (±8.67)	0.517 (±0.288)	0.476 (±0.312)
	Adj-Rank	0.452 (±0.247)	50.50% (±28.44)	65.57% (±23.74)	79.78% (±16.89)	94.32% (±8.20)	0.525 (±0.264)	0.462 (±0.345)
	Adj-Soft-Vote	**0.463 (±0.258)**	51.63% (±29.54)	65.85% (±24.61)	80.67% (±16.54)	95.68% (±6.65)	0.589 (±0.249)	0.531 (±0.301)

[1] The highest performances in term of each metric (column) in each setting (DTA and Intervention) are highlighted in bold fonts.

Multiagent debate

The debating strategy showed model-specific benefits. For most primary QA models in both the DTA and Intervention settings, MAD improved over QA baselines on most performance metrics except for GPT-MAD. GPT-MAD's performances on DTA slightly dropped in terms of MAP and R@5%, while significantly outperforming GPT in term of WSS. On Intervention, GPT-MAD had more mixed performances. It recorded a slight drop in MAP and WSS, but performed slightly better on recall. Comparatively, Gemini-MAD's performance gain over its QA counterpart is much more stable. Across DTA and Intervention, Gemini-MAD outperformed Gemini on most performance metrics such as MAP, WSS@95% and WSS@100%. Although the performance gains of Gemini-MAD cannot be said significant, they are not marginal either. Also note that, Gemini is the strongest QA model, outperforming other QA baselines by large margins.

Notably and surprisingly, it was the weakest model Haiku which benefited most from MAD. On DTA, Haiku-MAD's MAP and WSS@95% increased from 0.182 and 0.407 to 0.286 and 0.540, respectively, while on Intervention from 0.290 and 0.430 to 0.419 and 0.599. It is worth noting that, on Intervention, Haiku-MAD was overall the second best-performing model and was the best in term of R@5% and WSS@100%. In summary, the debating models, except GPT-MAD, exhibited good performances almost on par with the much more expensive adjudication methods, which highlights a promising future of MAD systems in SR automation.

LLM-based adjudication

Between LLM-based adjudication approaches, the adjudicator-as-a-ranker method (Adj-Rank) outperformed adjudicator-as-a-judge (Adj-Judge) in most important metrics like MAP, R@50% and WSS@95%, demonstrating that *preserving diverse perspectives through weighted averaging* is likely more effective than selecting a single "best" answer. Although adjudication did not beat Soft-Vote, it was much stronger than most other competitors in both settings except that both adjudicator variants were outperformed by Haiku-MAD on Intervention. These results not only underscore the potential value of the increasingly popular "LLM-as-a-Judge" paradigm [49] and the superiority of adjudicator-as-a-ranker as a novel contribution to this paradigm, but also highlight the high potential of the multiagent debating paradigm, especially when cost-effectiveness is considered.

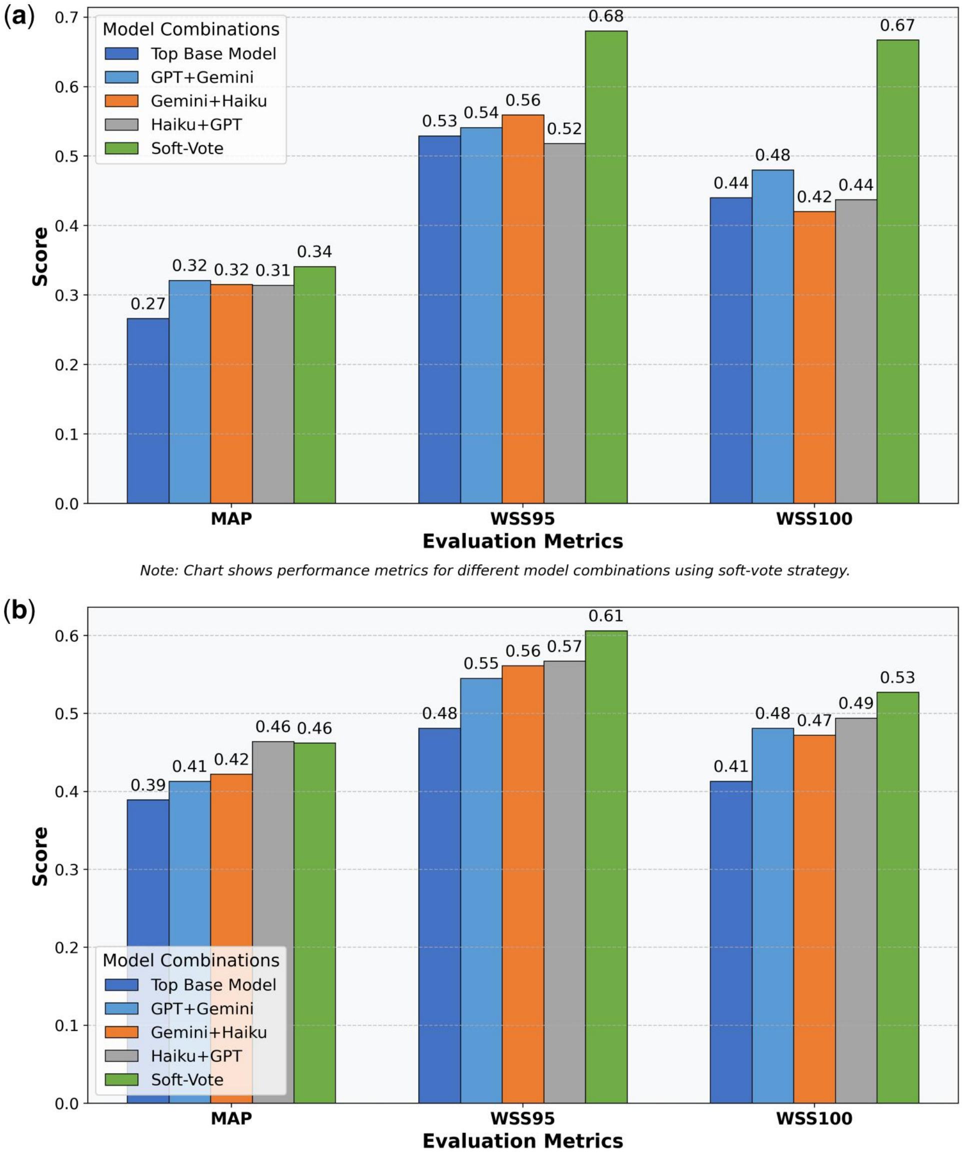
Peer-review investigation

Individual review performance varied substantially. Comprehensive per-review breakdowns showing all metrics for each SR are provided in Supplementary Files 2 and 3, enabling readers to assess method performance under conditions similar to their specific review characteristics. Multi-LLM collaboration showed greater benefits for reviews with moderate inclusion rates (5%–15%) and higher inter-model disagreement, as evidenced by the ablation study in Fig. 1 and the correlation analysis in Fig. 2. Soft-Vote achieved consistent improvements across all 28 reviews (MAP range: 0.182–0.687.), while benefits from multiagent debate were most pronounced for weaker models (Haiku-MAD: +0.129 MAP improvement on Intervention, $P < 0.001$). The bootstrap confidence intervals of the performances of each method are presented in Supplementary File 4. Comprehensive statistical comparisons, including effect sizes and bootstrap confidence intervals, are provided in the Supplementary File 5.

Insights, observed strengths and weaknesses

Critics of majority voting

The majority voting strategy Soft-Vote demonstrates consistent superiority (Table 1). Figure 1 shows the results of ablation study of this simple ensemble approach where we see that any two-model combination also achieved substantial improvements over individual models. Particularly, GPT+Haiku performed surprisingly well, almost rivalling Soft-Vote on Intervention in some metrics, like MAP (0.46



Note: Chart shows performance metrics for different model combinations using soft-vote strategy.

Note: Chart shows performance metrics for different model combinations using soft-vote strategy.

Figure 1 Ablation study of majority voting. (a) DTA setting. (b) Intervention setting.

vs. 0.46), WSS@95% (0.57 vs. 0.61), and WSS@100% (0.49 vs. 0.53). These findings suggest that benefits from diversity begin with just two base models and increase as more models are added.

Correlation analysis in Fig. 2 reveals moderate correlations between the QA models (Spearman's rank correlations: 0.48–0.56 on DTA and 0.49–0.52 on intervention), indicating each primary model captures different aspects of relevance—a diversity that an ensemble approach can effectively leverage. `Soft-Vote` shows high correlation with each individual model (Spearman: 0.71–0.86 on DTA and 0.68–0.87 on intervention), suggesting it preserves their strengths while mitigating weaknesses. The central role of diversity will be discussed in more detail in the "Model Diversity" subsection.

The strong performance combined with computational efficiency (see the cost breakdown in Table 2 and Supplementary Appendix E in Supplementary File 1 for the detailed cost estimation), which is less than 1/14 of adjudication and at most around 1/186 of the cost of a single human reviewer. This positions Majority Voting as an excellent default choice for early-stage screening. The ablation results suggest that even resource-constrained implementations using just two diverse models can achieve significant benefits.

Critics of multiagent debate

The debating strategy aims to improve screening performance by leveraging collective intelligence through exchanging structured argument. Overall, the MAD results provide several useful insights for designing MAD systems. Multiagent debate indeed brings performance gains, even with agents that are much weaker. The rethinking process is obviously the key to the success of MAD, which arguably improves performance through three mechanisms: (i) error correction when agents encounter opposing viewpoints with supporting evidence, (ii) uncertainty reduction by considering multiple

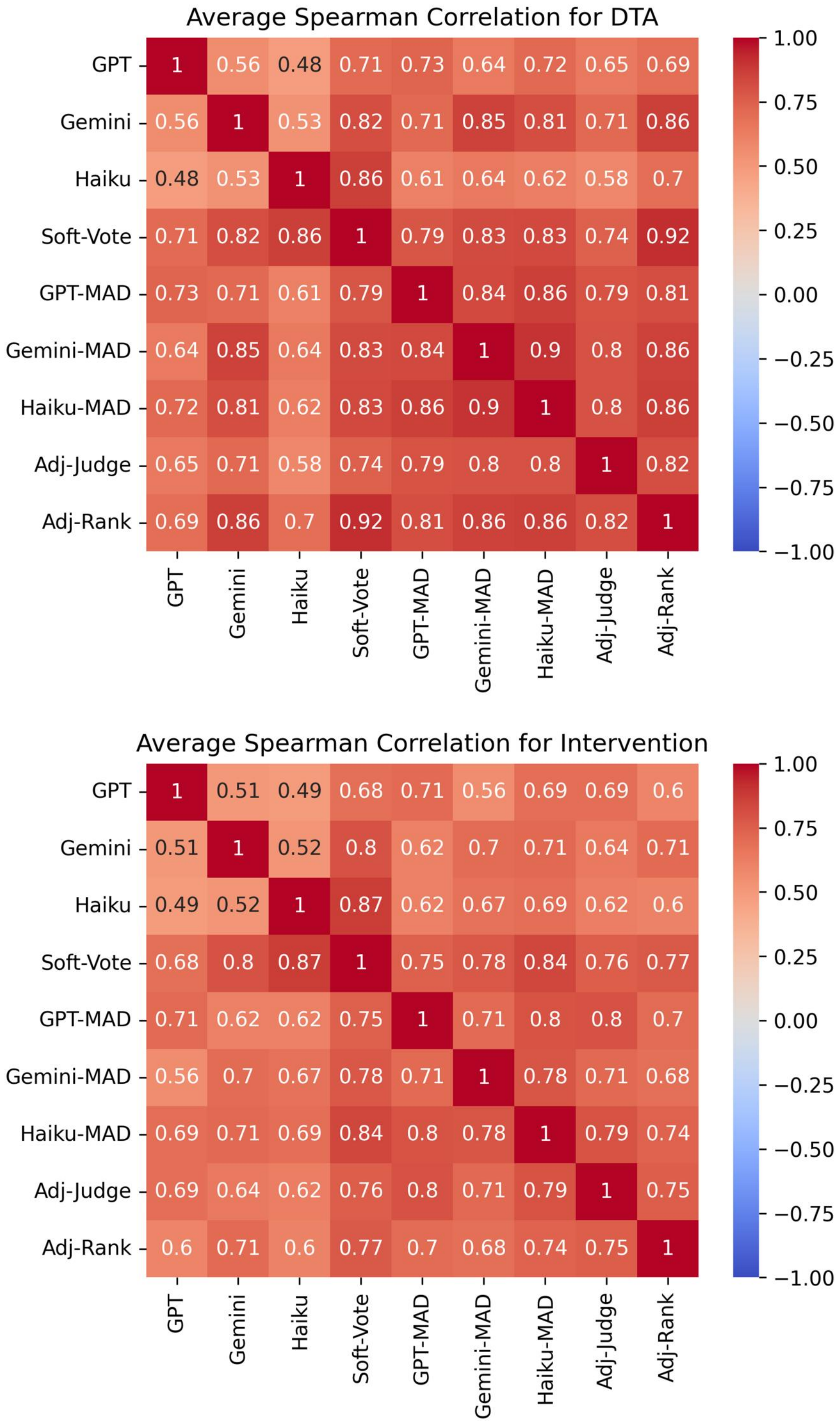


Figure 2 Correlation analysis between models. (a) DTA setting. (b) Intervention setting.

perspectives on ambiguous cases, and (iii) mitigation of model-specific bias through exposure to alternative interpretations.

Meanwhile, our analysis shows model-dependent outcomes as in Fig. 3. Benefits brought by MAD seem to be affected by agents' capabilities. While weaker agents will likely bring less benefits to stronger ones, they may benefit more from exposure to alternative perspectives. In reverse, opinions from stronger peers may have more significant impacts on self-reflection and decision-making. Both conform well with common sense. More specifically, `Gemini-MAD` showed minimal to negligible improvement over the QA baseline, possibly due to the latter's high performance, less influence from external arguments, or sub-optimal prompt tailoring. `GPT-MAD` exhibited more varied results. Despite obvious improvements on DTA, `GPT-MAD` stumbled at improving screening performance on Intervention across various metrics. Comparatively, `Haiku-MAD` obtained pronounce gains from debating. Although `Haiku` is notably much weaker than the other two competitors in medical abstract screening, it demonstrates a clear capacity for reasoning refinement through interaction with peers. This

Table 2 LLM pricing, cost breakdown, and runtime.

Setting	Type	Model	Input price	Input cost	Output price	Output cost	Total cost	Time (h)
DTA	Question answering	GPT-4o Mini	0.15	$2.54	0.6	$8.54	$11.08	100.6
		Gemini 1.5 Flash	0.075	$1.27	0.3	$4.17	$5.44	39.32
		Claude 3 Haiku	0.25	$4.24	1.25	$23.66	$27.90	62.5
	Majority voting	Soft-Vote	—	$8.05	—	$36.38	$44.43	202.42
	Debating	GPT-4o Mini	0.15	$18.32	0.6	$47.38	$65.70	334.52
		Gemini 1.5 Flash	0.075	$13.19	0.3	$42.64	$55.82	252.3
		Claude 3 Haiku	0.25	$25.16	1.25	$65.15	$90.31	280.26
	Voting on debating	MAD-Soft-Vote	—	$56.67	—	$155.17	$211.84	462.24
	Adjudication	Gemini 1.5 Pro	3.5	$263.49	10.5	$389.65	$653.15	906.32
Intervention	Question answering	GPT-4o Mini	0.15	$4.25	0.6	$12.56	$16.81	149.0
		Gemini 1.5 Flash	0.075	$2.12	0.3	$5.55	$7.67	58.27
		Claude 3 Haiku	0.25	$7.08	1.25	$34.65	$41.73	92.6
	Majority voting	Soft-Vote	—	$13.45	—	$52.76	$66.21	299.8
	Debating	GPT-4o Mini	0.15	$28.77	0.6	$68.74	$97.51	495.97
		Gemini 1.5 Flash	0.075	$21.11	0.3	$61.76	$82.88	373.92
		Claude 3 Haiku	0.25	$38.98	1.25	$94.45	$133.43	415.22
	Voting on debating	MAD-Soft-Vote	—	$88.86	—	$224.95	$313.81	685.3
	Adjudication	Gemini 1.5 Pro	3.5	$394.11	10.5	$565.80	$959.90	1342.8

phenomenon may have shed some *unusual light on the design of cost-effective LLM-based MAD systems* because stronger individual models may not always benefit most from multiagent debate.

The success of Haiku-MAD may to some extent lie in Haiku being more “diverse” from GPT (seen from Haiku’s comparatively low correlations with others: for example 0.48 and 0.53 on DTA in Fig. 2), but should be more rooted in its capability of leveraging peers’ wisdom. For example, although Gemini 1.5 Flash is the strongest individual model, it inclines to stick to its own decision, which can be demonstrated by the high correlation between the debating model Gemini-MAD and the QA baseline Gemini (e.g. 0.85 on DTA and 0.7 on intervention). GPT-4o Mini has the same but slightly weaker inclination. For instance, GPT-MAD is most correlated with GPT at 0.73 on DTA and 0.71 on Intervention. Notably and surprisingly, Claude 3 Haiku is the only debating model which perhaps takes more peer opinions than its own, which is implied from the fact that Haiku-MAD is more correlated with GPT and Gemini than with Haiku. The findings may have an important implication: *It might not be individual agent’s own capability but its capability of assimilating different opinions that makes multiagent debate systems work.*

The experiments also raise a question about whether MAD maintains model heterogeneity. Correlations among the three debating models have become much stronger than the correlations among the QA models, which is an expected phenomenon when peers gradually converge with the cohort while the cohort collectively improve. The ensemble over debating models, MAD-Soft-Vote, does not improve a lot over the best debating models (see Table 1), which further demonstrates the *core role of model heterogeneity* [58].

Critics of adjudication

Adjudication strategies introduce a hierarchical decision-making layer in a multi-LLM collaborative framework. Table 3 shows the average score assigned to each primary model by the adjudicator (here Gemini 1.5 Pro), along with how often each primary model was deemed the best or worst performing model by the adjudicator. The “Main Results: Screening Performance Enhancement” subsection shows Gemini (Gemini 1.5 Flash) is the strongest individual model. Indeed, the Gemini 1.5 Pro adjudicator has selected Gemini (Gemini 1.5 Flash) as the “best model” in over 50% and 45% of cases of DTA and Intervention, respectively, while its opinions about GPT (GPT-4o Mini) and Haiku (Claude 3 Haiku) were more or less balanced. Not surprisingly, Haiku was most frequently deemed the “worst model”. On Intervention, the chance of Haiku being rated as the worst model is significantly higher, which is likely linked to the fact that Haiku’s performance gap from GPT is also bigger. On the contrary, the chance of Gemini being rated as the best model drops possibly because Gemini's performance gain over other models on Intervention is less significant than on DTA. These findings have several implications. First, the adjudicator may indeed have some good capabilities in deciding the more appropriate primary model on a case by case basis, a key reason for significant screening performance improvement by both adjudication methods. In the meanwhile, they have also revealed *a potential bias of the adjudicator toward its own model family*, in corroboration with findings reported in Dietterich [59]. This is also reflected by the fact that the correlations between Gemini and the adjudicators is often stronger (Fig. 2), although more experiments are expected in the abstract screening context. This potential bias is a critical consideration apart from the significantly higher cost, as it could inadvertently reinforce existing preferential bias or limit the diversity of perspectives.

Model diversity

Model diversity plays a central role in ensemble performance [60]. The Soft-Vote approach, combining three distinct QA models, benefits from a “healthy” inter-model disagreement (i.e. model heterogeneity [58]), enabling it to balance strengths and mitigate weaknesses across agents. To make ensemble work, it is also important to make good trade-off between maximizing model accuracy and maintaining sufficient diversity [61]. MAD-Soft-Vote in Table 1 ensembles three debating models, each of which refines its reasoning through peer input before voting. While MAD improves individual model decisions, it reduces overall diversity among different debating models due to convergence in reasoning, demonstrated by the high correlations between GPT-MAD, Gemini-MAD and Haiku-

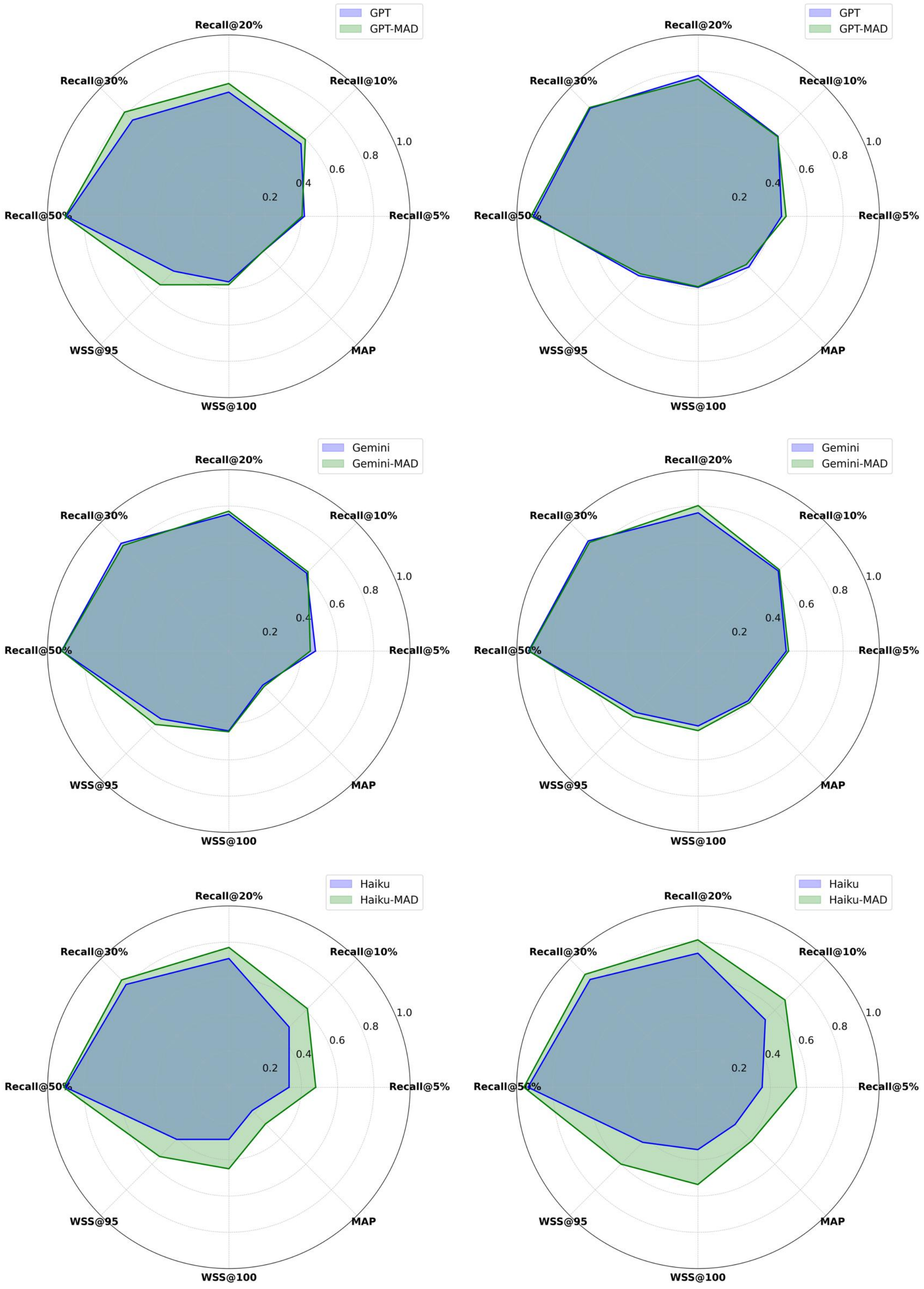


Figure 3 Multiagent debate versus QA models. (a) GPT-MAD versus GPT on DTA (b) GPT-MAD versus GPT on intervention. (c) Gemini-MAD versus Gemini on DTA (d) Gemini-MAD versus Gemini on intervention. (e) Haiku-MAD versus Haiku on DTA. (f) Haiku-MAD versus Haiku on intervention.

MAD. This leads to marginal gains of MAD-Soft-Vote over the best debating models and causes MAD-Soft-Vote to underperform Soft-Vote. Comparatively, Adj-Soft-Vote, which combines two adjudication models, gains significant improvement on Intervention, because of model diversity between the two adjudicators, demonstrated by the fact that the correlation between Adj-Judge and Adj-Rank on Intervention is much lower than that on DTA (see Fig. 2). In summary, while both debating and adjudication offer refinements, preserving model diversity through simple ensembling (`Soft-Vote`) remains the most cost-effective and robust strategy under resource constraints.

Additional results: re-ranking

Table 4 shows further performance improvement using both macro- and micro-level re-ranking (for simplicity of comparison, $\alpha = \beta = 0.5$). Table 4 also includes two important baselines from Akinseloyin et al. [30]. `GPT_Cos_Sim_Criteria` ranks candidate studies based on their semantic relevance with the selection criteria. `GPT_QA_Soft_Both_ReRank` is the best approach in Alison et al. [2] that integrates both macro- and micro-level re-ranking, thus comparable to our re-ranking variants.

Table 3 Ratings of QA models by the adjudicator.

Setting	Model	Avg rating	Best (%)	Worst (%)
DTA	GPT	0.855 (±0.018)	9.86	21.62
	Gemini	0.917 (±0.019)	51.92	10.17
	Haiku	0.762 (±0.021)	14.4	28.96
Intersection	GPT	0.887 (±0.016)	14.45	12.98
	Gemini	0.887 (±0.031)	45.15	19.33
	Haiku	0.76 (±0.086)	10.71	34.09

`Soft-Vote w/o re-ranking` significantly outperformed the best results in Akinseloyin et al. [30], particularly in WSS. Although GPT-3.5 was used in Akinseloyin et al. [30], we can still claim the benefit of multi-LLM collaboration based on the significant performance gain of `Soft-Vote w/o re-ranking` over the primary models. `GPT_Cos_Sim_Criteria's` decent performance highlights its plausible "model heterogeneity" that can be leveraged for improving our multi-LLM collaboration approaches through re-ranking. Indeed, `Soft-Vote w/re-ranking` achieved notable improvements over `Soft-Vote w/o re-ranking` on both DTA and Intervention, reaching new states of the art in term of WSS@95%. Notably, re-ranking also consistently improved the performances of other collaborative strategies by large margins, making them almost rival `Soft-Vote w/ re-ranking` on DTA.

Limitations

While this study demonstrates the effectiveness of multi-LLM collaboration for abstract screening, several limitations should be acknowledged. Additionally, while our experiments concatenated titles and abstracts for processing, the small subset of title-only records (approximately 2%–3% of documents in TAR2019) provided limited context compared to full title-and-abstract screening. This may have slightly affected classification validity for these records, though sensitivity analyses suggested minimal impact on overall performance metrics.

Table 4 Performance improvements by re-ranking.

	Model	MAP	R@5%	R@10%	R@20%	R@50%	WSS@95%	WSS@100%
DTA	GPT cos sim criteria [2]	0.271	47.7%	62.8%	78.2%	94.1%	60.0%	51.3%
	GPT QA soft both ReRank [2]	0.315	43.8%	59.3%	76.6%	94.1%	56.6%	50.6%
	Soft-vote w/o re-ranking	0.341 (±0.166)	46.92% (±26.98)	64.67% (±28.37)	79.31% (±25.67)	94.84% (±8.83)	0.680 (±0.228)	0.667 (±0.266)
	Soft-vote w/ re-ranking	0.360 (±0.139)	56.52% (±34.49)	72.32% (±30.90)	84.67% (±24.87)	96.59% (±8.08)	0.708 (±0.220)	0.664 (±0.260)
	GPT-MAD w/ re-ranking	0.348 (±0.142)	57.47% (±35.83)	71.36% (±30.66)	82.86% (±25.94)	96.91% (±8.17)	0.690 (±0.235)	0.648 (±0.261)
	Gemini-MAD w/ re-ranking	0.376 (±0.139)	55.84% (±36.07)	71.63% (±30.89)	83.17% (±26.93)	96.62% (±22.03)	0.704 (±0.216)	0.655 (±0.252)
	Haiku-MAD w/ re-ranking	0.368 (±0.147)	57.66% (±35.09)	71.62% (±29.53)	82.67% (±26.52)	96.33% (±9.81)	0.705 (±0.220)	0.654 (±0.245)
	Adj-judge w/ re-ranking	0.364 (±0.134)	57.40% (±34.88)	71.92% (±31.45)	82.93% (±26.70)	97.09% (±8.22)	0.697 (±0.221)	0.674 (±0.240)
Intervention	Adj-rank w/ re-ranking	0.378 (±0.135)	56.85% (±36.60)	71.41% (±31.85)	84.11% (±25.45)	96.59% (±8.08)	0.706 (±0.226)	0.667 (±0.256)
	GPT cos sim criteria [2]	0.271	40.1%	54.4%	72.2%	92.0%	55.2%	49.9%
	GPT QA soft both ReRank [2]	0.450	52.6%	69.7%	81.6%	95.9%	60.0%	52.6%
	Soft-vote w/o re-ranking	0.462 (±0.122)	53.15% (±29.31)	66.71% (±25.78)	83.41% (±15.74)	96.82% (±5.63)	0.606 (±0.219)	0.527 (±0.270)
	Soft-vote w/ re-ranking	0.470 (±0.248)	56.59% (±32.28)	71.76% (±26.17)	84.89% (±17.33)	97.87% (±2.64)	0.696 (±0.210)	0.636 (±0.29)
	GPT-MAD w/ re-ranking	0.447 (±0.254)	56.11% (±31.98)	70.06% (±27.80)	83.16% (±20.22)	97.48% (±5.57)	0.658 (±0.225)	0.609 (±0.287)
	Gemini-MAD w/ re-ranking	0.470 (±0.268)	55.52% (±32.54)	71.63% (±26.30)	84.52% (±18.77)	97.39% (±5.69)	0.673 (±0.226)	0.633 (±0.278)
	Haiku-MAD w/ re-ranking	0.459 (±0.248)	55.78% (±31.60)	70.17% (±27.63)	84.53% (±19.78)	98.01% (±4.23)	0.690 (±0.217)	0.643 (±0.292)
	Adj-judge w/ re-ranking	0.482 (±0.248)	58.08% (±32.79)	72.05% (±26.33)	84.55% (±18.03)	97.57% (±5.32)	0.663 (±0.222)	0.616 (±0.290)

Second, our evaluation is confined to the biomedical TAR2019 benchmark, which comprises Cochrane SRs in clinical intervention and diagnostic technology assessment. Medical abstracts typically follow structured formats (e.g. IMRAD: Introduction, Methods, Results, and Discussion) with consistent terminology and well-established reporting standards such as CONSORT (Consolidated Standards of Reporting Trials) for trials and STARD (Standards for Reporting of Diagnostic Accuracy Studies) for diagnostic accuracy studies. These characteristics may make biomedical abstracts particularly amenable to LLM-based classification. The performance of our multi-LLM collaboration strategies in other domains such as social sciences, environmental studies, education research, or humanities where abstracts may be less structured, terminology more varied, and reporting standards less uniform, remains to be validated. Cross-domain evaluation represents an important direction for future work to establish the generalizability of these approaches.

Conclusion

This study presents an in-depth investigation into LLM-based multi-agent collaborative strategies for automating abstract screening in SRs. We successfully developed and evaluated three collaborative strategies—majority voting, multiagent debate and LLM-based adjudication, and demonstrated that collaboration among multiple, cost-effective LLMs has high potential to substantially reduce screening workload and cost. The collaborative frameworks effectively mitigate individual LLMs' biases through collective intelligence.

Majority voting emerged as the most robust solution, achieving consistent and significant performance gains in all settings. Analysis demonstrated the core role of model diversity (i.e. model heterogeneity) on the success of aggregating relatively weaker screening models. While debating improves screening performance, its benefits are more model-specific. Models that stick to their own decisions are less likely benefiting from peers, which is deemed an important insight for developing effective multiagent debating systems in the context of abstract screening. Nevertheless, we argue that it is worth conducting further research in this strategy with cost-effective lightweight LLMs, especially when taking into consideration its strikingly lower costs than the recent LLM-as-a-Judge paradigm.

The economic implications are also substantial. Majority voting only costs less than 1/14 of that of adjudication methods based on a strong LLM, and achieves more than 186× cost reduction compared to single human reviewer based on a conservative estimation according to British academic salary scales. By demonstrating that multi-LLM collaboration can achieve superior performance at a substantially lower cost, this research offers a pathway toward making SR automation both more effective and affordable.

Author contributions

Opeoluwa Akinseloyin (Conceptualization [equal], Formal analysis [equal], Investigation [lead], Methodology [equal], Software [equal], Validation [lead], Visualization [equal], Writing—original draft [equal]), Xiaorui Jiang (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Vasile Palade (Investigation [supporting], Methodology [supporting], Project administration [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary material

Supplementary material is available at *Biology Methods and Protocols* online.

Conflicts of interest

None declared.

Funding

None declared.

Data availability

All data of analysis are published in the main text and the supplementary materials, while all experimental configurations, including model versions and temperature settings, are detailed in the main text. Regarding the raw data for the experiments, the TAR2019 benchmark dataset is available in the CLEF-TAR repository at https://github.com/CLEF-TAR/tar/tree/master/2019-TAR (Task 2 of CLEF eHealth 2019 Technology-Assisted Review). The titles and abstracts for all documents in the TAR2019 dataset are copyrighted but can be extracted from PubMed programmatically via the PubMed API using the provided PubMed IDs (PMIDs) in the CLEF-TAR repository. The selection criteria for each SR are included in the CLEF-TAR repository. The converted inclusion criteria questions used in our QA framework are provided in Supplementary Appendix D of Supplementary File 1. To enhance reproducibility, we have created a public GitHub repository containing the prompts used for all collaboration strategies along with implementation details: https://github.com/Ope-Akinseloyin/Multi_LLM-Citation-Screening.

References

1. Tsafnat G, Glasziou P, Choong MK *et al.* Systematic review automation technologies. *Syst Rev* 2014;**3**:74–15. https://doi.org/10.1186/2046-4053-3-74
2. Alison O-E, Thomas J, McNaught J *et al.* Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015;**4**:5. https://doi.org/10.1186/2046-4053-4-5
3. Borah R, Brown AW, Capers PL *et al.* Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017;**7**:e012545. https://doi.org/10.1136/bmjopen-2016-012545
4. Rathbone J, Carter M, Hoffmann T *et al.* Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. *Syst Rev* 2015;**4**:6. https://doi.org/10.1186/2046-4053-4-6
5. Bramer WM, Rethlefsen ML, Kleijnen J *et al.* Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev* 2017;**6**:245. https://doi.org/10.1186/s13643-017-0644-y
6. Edwards P, Clarke M, DiGuiseppi C *et al.* Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Stat Med* 2002;**21**:1635–40. https://doi.org/10.1002/sim.1190
7. Cohen AM, Hersh WR, Peterson K *et al.* Reducing workload in systematic review preparation using automated citation

classification. *J Am Med Inform Assoc* 2006;**13**:206–19. https://doi.org/10.1197/jamia.M1929

8. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. *J Am Med Inform Assoc* 2009;**16**:690–704. https://doi.org/10.1197/jamia.M3162
9. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. *Res Synth Methods* 2011;**2**:1–14. https://doi.org/10.1002/jrsm.27
10. Van der Mierden S, Tsaioun K, Bleich A *et al.* Software tools for literature screening in systematic reviews in biomedical research. *ALTEX* 2019;**36**:508–17. https://doi.org/10.14573/altex.1902131
11. Harrison H, Griffin SJ, Kuhn I *et al.* Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Med Res Methodol* 2020;**20**:7. https://doi.org/10.1186/s12874-020-0897-3
12. Wallace BC, Trikalinos TA, Lau J *et al.* Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 2010;**11**:55–11. https://doi.org/10.1186/1471-2105-11-55
13. Hughes M, Li I, Kotoulas S *et al.* Medical text classification using convolutional neural networks. In: Randell R, Cornet R, McCowan C, Peek N, and Scott PJ (eds.), *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, IOS Press, 2017, 246–50.
14. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *IDA* 2002;**6**:429–49. https://doi.org/10.3233/IDA-2002-6504
15. Garcia EA, Haibo H. Learning from imbalanced data. *IEEE Trans Knowledge Data Eng* 2009;**21**:1263–84. https://doi.org/10.1109/TKDE.2008.239
16. Wallace BC, Small K, Brodley CE *et al.* Active learning for biomedical citation screening. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, pp. 173–82. New York, NY, USA: Association for Computing Machinery, 2010. https://doi.org/10.1145/1835804.1835829
17. van de Schoot R, de Bruin J, Schram R *et al.* An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell* 2021;**3**:125–33. https://doi.org/10.1038/s42256-020-00287-7
18. Miwa M, Thomas J, O'Mara-Eves A *et al.* Reducing systematic review workload through certainty-based screening. *J Biomed Inform* 2014;**51**:242–53. https://doi.org/10.1016/j.jbi.2014.06.005
19. Michiel PB, Greijn B, Messina Coimbra B *et al.* Combining large language model classifications and active learning for improved technology-assisted review. In: Bunse M, Herde M, Krempl G, Lemaire V, Tharwat A, Tuan Pham M, and Saadallah A (eds.), *Proceedings of the Workshop on Interactive Adaptive Learning co-Located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2024)*, Vilnius, Lithuania, volume 3770 of CEUR Workshop Proceedings, 2024, pp. 77–95. https://ceur-ws.org/Vol-3770/paper8.pdf
20. Gordon VC, Grossman MR. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv: 1504.06868*, 2015.
21. Olofsson H, Brolund A, Hellberg C *et al.* Can abstract screening workload be reduced using text mining? user experiences of the tool Rayyan. *Res Synth Methods* 2017;**8**:275–80. https://doi.org/10.1002/jrsm.1237
22. Brown T, Mann B, Ryder N *et al.* Language models are few-shot learners. *Adv Neural Inform Proces Syst* 2020;**33**:1877–901.
23. Liu P, Yuan W, Fu J *et al.* Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 2023;**55**:1–35. https://doi.org/10.1145/3560815
24. Wei J, Wang X, Schuurmans D *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inform Process Syst* 2022;**35**:24824–37.
25. Lewis P, Perez E, Piktus A *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inform Process Syst* 2020;**33**:9459–74.
26. Devlin J, Chang M-W, Lee K *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding, In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423
27. Radford A, Wu J, Child R *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* 2019;**1**:9.
28. Wang S, Scells H, Koopman B *et al.* Can ChatGPT write a good Boolean query for systematic review literature search? In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1426–36. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3539618.3591703
29. Guo E, Gupta M, Deng J *et al.* Automated paper screening for clinical reviews using large language models. *J Med Internet Res* 2024;**26**:e48996. https://doi.org/10.2196/48996
30. Akinseloyin O, Jiang X, Palade V. A question-answering framework for automated abstract screening using large language models. *J Am Med Inform Assoc* 2024;**31**:1939–52. https://doi.org/10.1093/jamia/ocae166
31. Landschaft A, Antweiler D, Mackay S *et al.* Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. *Int J Med Inform* 2024;**189**:105531. https://doi.org/10.1016/j.ijmedinf.2024.105531
32. Vallamchetla SK, Abdelkader O, Elnaggar A *et al.* Do it faster with PICOS: generative AI-assisted systematic review screening. *J Biomed Inform* 2025;**168**:104860. https://doi.org/10.1016/j.jbi.2025.104860
33. Cao C, Arora R, Cento P, *et al.* Automation of systematic reviews with large language models. *medRxiv*, 2025. https://doi.org/10.1101/2025.06.13.25329541
34. Akinseloyin O, Jiang X, Paladel V. Weakly supervised active learning for abstract screening leveraging LLM-based pseudo-labeling. *medRxiv*, 2025. https://doi.org/10.1101/2025.08.24.25334314
35. Bender EM, Gebru T, McMillan-Major A *et al.* On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–23. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922
36. Gallegos IO, Rossi RA, Barrow J *et al.* Bias and fairness in large language models: a survey. *Comput Linguist* 2024;**50**:1097–179. https://doi.org/10.1162/coli_a_00524
37. Achiam J, Adler S, Agarwal S *et al.* Gpt-4 technical report. *arXiv Preprint arXiv: 2303.08774* 2023.
38. Oami T, Okada Y, Nakada T-A. Optimal large language models to screen citations for systematic reviews. *Res Synth Methods* 2025;**16**:859–75. https://doi.org/10.1017/rsm.2025.10014
39. Guo T, Chen X, Wang Y *et al.* Large language model based multi-agents: a survey of progress and challenges. In: *Proceedings of the*

Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp. 8048–57. International Joint Conferences on Artificial Intelligence Organization, 2024. Survey Track. https://doi.org/10.24963/ijcai.2024/890

40. Wang L, Ma C, Feng X *et al.* A survey on large language model based autonomous agents. *Front Comput Sci* 2024;**18**:186345. https://doi.org/10.1007/s11704-024-40231-1
41. Xi Z, Chen W, Guo X *et al.* The rise and potential of large language model based agents: a survey. *Sci China Inf Sci* 2025;**68**:121101. https://doi.org/10.1007/s11432-024-4222-0
42. Sanghera R, Thirunavukarasu AJ, El Khoury M *et al.* High-performance automated abstract screening with large language model ensembles. *J Am Med Inform Assoc* 2025;**32**:893–904. https://doi.org/10.1093/jamia/ocaf050
43. Zhang Z, Momeni Nezhad MJ, Gupta P *et al.* Enhancing ai for citation screening in literature reviews: improving accuracy with ensemble models. *Int J Med Inform* 2025;**203**:106035.
44. Li X, Wang S, Zeng S *et al.* A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 2024; **1**:9. https://doi.org/10.1007/s44336-024-00009-2
45. Liang T, He Z, Jiao W, *et al.* Encouraging divergent thinking in large language models through multi-agent debate. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–904, Miami, Florida: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.992
46. Du Y, Li S, Torralba A *et al.* Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv: 2305.14325*, 2023.
47. Chan C-M, Chen W, Su Y *et al.* ChatEval: Towards better LLM-based evaluators through multi-agent debate. In: *Proceedings of the Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria: OpenReview.net, 2024.
48. Yin Z, Sun Q, Chang C *et al.* Exchange- of-thought: Enhancing large language model capabilities through cross-model communication. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15135–53, Singapore: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.936
49. Li D, Jiang B, Huang L *et al.* From generation to judgment: opportunities and challenges of LLM-as-a-judge. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–91, Suzhou: Association for Computational Linguistics, 2025. https://doi.org/10.18653/v1/2025.emnlp-main.138
50. Gu J, Jiang X, Shi Z *et al.* Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. *arXiv preprint arXiv: 2411.15594*, 2025.
51. Chen X, Yi H, You M *et al.* Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ Digit Med* 2025;**8**:159. https://doi.org/10.1038/s41746-025-01550-0
52. Tang X, Zou A, Zhang Z *et al.* MedAgents: Large language models as collaborators for zero-shot medical reasoning. In: *Findings of the Association for Computational Linguistics*, pp. 599–621, Bangkok, Thailand: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.findings-acl.33
53. Lu M, Ho B, Ren D *et al.* TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5747–64, Miami, Florida: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.findings-emnlp.329
54. Lewis M, Liu Y, Goyal N *et al.* BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–80. Kerrville, TX, USA: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.703
55. Wu Q, Bansal G, Zhang J *et al.* AutoGen: enabling next-gen LLM applications via multi-agent conversation. In: *Proceedings of ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2023. https://openreview.net/pdf? id=uAjxFFing2
56. Shinn N, Cassano F, Gopinath A *et al.* Reflexion: language agents with verbal reinforcement learning. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, and Levine S (eds.), *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2023, 8634–52.
57. Chiang C-H, Lee H-V. Can large language models be an alternative to human evaluations? In: Anna R, Jordan B-G, and Naoaki O (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–31, Toronto, Canada: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.acl-long.870
58. Panickssery A, Bowman SR, Feng S. LLM evaluators recognize and favor their own generations. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, and Zhang C (eds.), *Advances in Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates, Inc., 2024, 68772–802.
59. Dietterich TG. Ensemble methods in machine learning. In: Kittler J, Roli F (ed.), *Multiple Classifier Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, 1–15.
60. Zhang H, Cui Z, Chen J. *et al.* Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv: 2502.08788*, 2025.
61. Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag* 2006;**6**:21–45. https://doi.org/10.1109/MCAS.2006.1688199

Biology Methods and Protocols, 2026, 11, 1–13
https://doi.org/10.1093/biomethods/bpag006
Methods Article