Journal Title Here, 2022, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

A LLM-based Multi-Agent Collaborative Approach for Screening Prioritization towards Automated Systematic Reviews

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

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

[*]Xiaorui Jiang. xiaorui.jiang@sheffield.ac.uk

FOR PUBLISHER ONLY Received on Date 2025; revised on Date Month Year; accepted on Date Month Year

Abstract

Objective: Systematic reviews (SRs) are essential for evidence-based practice but remain labor-intensive, especially during title and abstract screening. This study evaluates whether multiple large language model (multi-LLM) collaboration can improve screening prioritization while reducing costs.

Methods: Abstract screening was framed as a question-answering (QA) task using cost-effective LLMs. Three multi-LLM collaboration strategies were proposed and evaluated, including majority voting by averaging opinions of peers, multi-agent debate (MAD) for answer refinement, and LLM-based adjudication against answers of individual QA baselines. These strategies were evaluated on the CLEF eHealth 2019 Technology-Assisted Review benchmark using standard performance metrics in the domain, including Mean Average Precision (MAP), Recall@k%, and Work Saved over Sampling (WSS).

Results: Multi-LLM collaboration strategies significantly outperformed QA baselines. Majority voting was the most outstanding, achieving the highest MAP 0.4621 and 0.3409 on the subsets of SRs about clinical intervention and diagnostic technology assessment, respectively, with WSS@95% 0.6064 and 0.6798, enabling in theory up to 68% workload reduction at 95% recall of all included studies. MAD improved weaker models the most. The adjudicator-as-a-ranker method surpassed adjudicator-as-a-judge and was the second strongest approach, but with a significantly higher cost than majority voting and debating.

Conclusion: Multi-LLM collaboration can substantially improve abstract screening efficiency, and the success lies in model diversity. Making the the best use of diversity, majority voting stands out in terms of both fantastic performance and low cost compared to adjudication.Despite context-dependent gains and diminishing model diversity, MAD is still a cost-effective strategy and a potential direction of further research.

Key words: Systematic Review, Abstract Screening, Large Language Model, Multi-Agent System, Evidence Synthesis

Background and Significance

Systematic reviews serve as cornerstones for evidence-based practice, particularly in medicine and healthcare, by providing rigorous summaries of existing knowledge [46]. However, conducting SRs is notoriously labor-intensive, often requiring researchers to spend over several months [37, 4]. A primary bottleneck is the title and abstract screening phase, where the sheer volume of retrieved studies—sometimes as high as tens of thousands [40, 5]–presents formidable challenges amplified by the standard practice of involving multiple human annotators [17].

To tackle this grand challenge, there has been two decades of efforts on using AI to assist the screening task since Cohen et al.'s seminal works [13, 12]. Early automation efforts leveraged traditional machine learning techniques [37, 45], later progressing to deep learning models [49, 23]. These approaches faced significant challenges, such as the requirement for extensive labelled data, and the need for model retraining for each new review [37, 51, 25, 26, 24]. Active learning sought to alleviate these issues through iterative human feedback [50, 48, 42, 35], but faced challenges in achieving a high degree of automation due to the zero-shot nature of this problem [14, 36].

The advent of Large Language Models (LLMs) marked a paradigm shift, offering unprecedented zero-shot learning capabilities for tasks like abstract screening [6, 32, 54, 28, 15, 39]. Recent studies have demonstrated LLMs' potential across various SR stages, from search query generation [53] to literature screening tasks [20, 2, 47] and even the whole SR pipeline [7]. However, single LLMs inevitably suffer inherent model biases [3, 18] and weak alignment with

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

nuanced human judgment [1]. These limitations have catalyzed interest in enhancing reliability through multiple large language model (multi-LLM) collaboration, aka LLM-based Multi-Agent Systems (MAS) [21, 52, 56]. For example, ensemble approaches, a primitive form of multi-model collaboration, have demonstrated superior performance for abstract screening by combing the decisions from a range of LLMs [41, 59].

LLM-based MAS (hear agent is an LLM) has emerged as a powerful paradigm across various domains, with extensive research demonstrating its effectiveness [22, 30, 33]. Key collaboration strategies include debating mechanisms [16, 8], where agents engage in structured argumentation [16, 57, 31], and adjudication approaches leveraging, for example, the LLM-as-a-Judge frameworks [29, 19]. These approaches aim to mitigate individual model weaknesses through collection reasoning among peers and have been demonstrated effective for complex reasoning tasks in medical domains [10, 44, 34]. Yet, a comprehensive investigation of multi-LLM collaboration hasn't been presented in the context of abstract screening. This paper makes a comprehensive exploration of multi-LLM collaboration towards automated abstract screening. Our contributions are threefold: (1) We investigate multi-LLM collaborative strategies, including ensembling, debating, and adjudication, towards systematic review automation. (2) We comparatively evaluate different strategies to identify the most robust approaches. (3) We empirically analyze the core success factors that enable collaborative reasoning processes to mitigate errors of individual LLMs.

Methodology

LLM-based Question Answering for Screening

Consistent with [2], we formulate the abstract screening task using a question-answering (QA) framework. Suppose each SR constitutes an unannoated dataset of (the titles and abstracts of) candidate studies (i.e., *documents*) $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, where N is the total number of documents and d_i 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, $\mathcal{Q} = \{q_1, \ldots, q_K\}$, that every included study must satisfy. Given a document d_i ($i = 1, ..., N$), each inclusion criteria question q_k ($k = 1...K$) will be answered by an LLM-based QA model $\mathbf{M}$, with the answer $a_{i,k}^{\mathbf{M}}$ 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 presented in the Supplementary Materials:

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

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

Given a QA model $\mathbf{M}^{\mathtt{qa}}$, we use the same method in [2] 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 $a_{i,k}^{\mathbf{M}^{\mathtt{qa}}}$ and $r_{i,k}^{\mathbf{M}^{\mathtt{qa}}}$) has a positive sentiment, according to a pretrained BART model [27]:

$$\text{score}(d_i, q_k; \mathbf{M}^{\mathtt{qa}}) = \text{Prob}_{\mathtt{BART}}(\text{Positive}|a_{i,k}^{\mathbf{M}^{\mathtt{qa}}}, r_{i,k}^{\mathbf{M}^{\mathtt{qa}}}). \quad (1)$$

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

$$\text{score}(d_i, \mathcal{Q}; \mathbf{M}^{\mathtt{qa}}) = \sum_{k=1}^{K} \text{score}(d_i, q_k; \mathbf{M}^{\mathtt{qa}}). \quad (2)$$

Multi-Agent Collaboration for Screening

Inspired by recent work in LLM-based MAS [21, 16, 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 the Supplementary Materials.

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 d_i, the soft voting score is defined as the average of the scores for each primary model:

$$\text{score}_{\mathtt{vote}}(d_i, q_k; \mathcal{M}) = \frac{1}{L} \sum_{l=1}^{L} \text{score}_{\mathtt{vote}}(d_i, q_k; \mathbf{M}_l). \quad (3)$$

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

$$\text{score}_{\mathtt{vote}}(d_i, \mathcal{Q}; \mathcal{M}) = \sum_{k=1}^{K} \text{score}_{\mathtt{vote}}(d_i, q_k; \mathcal{M}). \quad (4)$$

Multi-Agent Debate

This strategy introduces a step of cross-agent communication, reflection, and refinement[43, 16]. After an initial round of independent QA, each agent is presented with the answers and reasoning of other agents for the same 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:

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

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

$$\mathcal{I}_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}} =< \mathcal{I}_{i,k}^{\mathtt{qa}}, \mathcal{O}_{i,k}^{\mathbf{M}_l^{\mathtt{qa}}}, \mathcal{O}_{i,k}^{\mathbf{M}_{l'}^{\mathtt{qa}}}|_{l'=1,l'\neq l}^{L} > ,$$

and the output of the debating model is as follows

$$\mathcal{O}_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}} =< a_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}, r_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}, v_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}, c_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}, e_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}} > ,$$

with $a_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}$, $r_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}$ and $v_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}$ being the new *a*nswer, the *r*easoning, and the confidence *v*alue of the debating agent on the new answer, $c_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}$ indicating whether the debating agent *c*hanges the original answer ("Yes" or "No"), and $e_{i,k}^{\mathbf{M}_l^{\mathtt{deb}}}$ holding the *e*xplanation for why the answer is changed or kept the same. Note that $\mathbf{M}_l^{\mathtt{deb}}$ and $\mathbf{M}_l^{\mathtt{qa}}$ refer to the same model $\mathbf{M}_l \in \mathcal{M}$ used for different purposes, taking different inputs and generating different outputs.

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:

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

where

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

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.

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 [11] . 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:

$$\mathcal{O}_{i,k}^{\mathbf{M}^{\text{adj}}} = \mathbf{M}^{\text{adj}}(\mathcal{I}_{i,k}^{\mathbf{M}^{\text{adj}}}), \quad (7)$$

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,

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

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 $g^{\mathbf{M}_l^{\text{qa}}}$ being the *g*rading of a QA model (a rating value between 0 and 1), and $b_{i,k}^{\mathbf{M}^{\text{deb}}}$ 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" [29], or called "meta-agent" in other literature [9]. Similar to other methods, the score of a document d_i with respect to each criteria question q_k by an adjudicator as a judge is the probability that its answer text (i.e., $a_{i,k}^{\mathbf{M}^{\text{adj}}}$) has a positive sentiment based on BART:

$$\text{score}_{\texttt{judge}}(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}) = \text{Prob}_{\texttt{BART}}(\text{Positive}|a_{i,k}^{\mathbf{M}^{\text{adj}}}), \quad (8)$$

and the final score of each document is:

$$\text{score}_{\texttt{judge}}(d_i, \mathcal{Q}; \mathbf{M}^{\text{adj}}, \mathcal{M}) = \sum_{k=1}^{K} \text{score}_{\texttt{judge}}(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}) \quad (9)$$

Adjudicator as a Ranker. Alternatively, we ask the adjudicator LLM to rate the quality of individual answers and generate a grade for each primary model on a document-question pair, denoted by $g_{i,k}^{\mathbf{M}_l^{\text{qa}}}$ $(l = 1, \ldots, L)$, 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:

$$\text{score}_{\texttt{rank}}(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}) = \frac{1}{L} \sum_{l=1}^{L} g_{i,k}^{\mathbf{M}_l^{\text{adj}}} \cdot \text{score}(d_i, q_k; \mathbf{M}_l^{\text{qa}}). \quad (10)$$

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

$$\text{score}_{\texttt{rank}}(d_i, \mathcal{Q}; \mathbf{M}^{\text{adj}}, \mathcal{M}) = \sum_{k=1}^{K} \text{score}_{\texttt{rank}}(d_i, q_k; \mathbf{M}^{\text{adj}}, \mathcal{M}) \quad (11)$$

Re-ranking

Screening prioritization performance can be further improved through re-ranking as in [2]. 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 $\text{rel}(d_i, \mathcal{Q})$. A macro-level re-ranking score is calculated as follows:

$$\text{score}^*_{\texttt{rr-mac}}(d_i, \mathcal{Q}) = (1 - \alpha) \cdot \text{score}(d_i, \mathcal{Q}) + \alpha \cdot \text{rel}(d_i, \mathcal{Q}), \quad (12)$$

where $\alpha \in (0, 1)$, and $\text{score}(di, \mathcal{Q})$ is the score of d_i that is calculated according to either Eq. (4), Eq. (5), Eq. (9) or Eq. (11).

***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 $\text{rel}(d_i, q_k)$. Micro-level re-ranking first calculates a new score for each document with respect to each question as follows,

$$\text{score}^*_{\texttt{rr-mic}}(d_i, q_k) = (1 - \beta) \cdot \text{score}(d_i, q_k) + \beta \cdot \text{rel}(d_i, q_k), \quad (13)$$

where $\beta \in (0, 1)$ and $\text{score}(d_i, q_k)$ is calculated according to either Eq. (3), Eq. (6), Eq. (8) or Eq. (10). Then, a new score for each document is calculated by summing up the scores with respect to all criteria questions:

$$\text{score}^*_{\texttt{rr-mic}}(d_i, \mathcal{Q}) = \sum_{k=1}^{K} \text{score}^*_{\texttt{rr-mic}}(d_i, q_k). \quad (14)$$

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 screening prioritization methods in systematic reviews, including 20 Cochrane reviews about clinical intervention (Intervention; in total 39,792 documents) and 8 reviews about diagnostic technology assessment (DTA;26,830). The dataset is detailed in the Supplementary Materials.

We employ standard TAR evaluation metrics, including the rank of the *L*ast *Rel*evant document (L_{Rel}), *M*ean *A*verage *P*recision (MAP), *R*ecall *at* top k% of documents screened ($R@k$%, for $k = 5, 10, 20, \ldots, 60$), and *W*ork *S*aved over *S*ampling (WSS) *at* the recall level of R% (WSS@R%, for $R =$

95, 100). WSS@R% measures the proportion of the collection that can be skipped when R% of the relevant documents are retrieved [13]. Metrics are calculated per SR and then averaged across all the SRs in each of the two categories.

Large Language Models 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 Eq. (2), named by `GPT`, `Haiku`, and `Gemini`;
- The soft voter according to Eq. (4), named by `Soft-Vote`;
- The three multi-agent debate models according to Eq. (5)), named by `GPT-MAD`, `Haiku-MAD`, and `Gemini-MAD`;
- The LLM adjudicator as a judge (`Adj-Judge`) and as a ranker (`Adj-Rank`) according to Eq. (9) and Eq. (11) respectively.

We also compare the re-ranking variants of the aforementioned approaches that integrate both micro- and macro-level re-ranking, according to Eqs. (14) and (12). The GPT embedding model "text- embedding-ada-002" is used for generate text embeddings.

Results and Discussion

Screening Performance Enhancement

Table 1 presents the performance comparisons. On both DTA and Intervention, `Soft-Vote` substantially outperformed the individual baseline models, and obtained the highest MAP, recalls and WSS values among all competitors. R@10% reached 0.6467 (±0.3233) and 0.6671 (±0.1679) 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 0.6798 (±0.2389) WSS@95% and 0.6668 (±0.2786) WSS@100% on DTA, and 0.6064 (±0.1400) WSS@95% and 0.5274 (±0.1727)) WSS100%. Considering its relatively low cost, the simple approach `Soft-Vote` is both strong and cost-effective (refer to Sect. **??** for detailed analysis). The simple aggregation effectively mitigates individual model weaknesses and biases. The theoretical average workload reductions can be as high as 67.98% and 60.64% on the 8 DTA and 20 Intervention SRs, respectively, when the recall threshold of relevant studies is 95%, which is a widely-adopted critical requirement for avoiding causing significant damage to the reliability of the conclusions of an SR.

The debating strategy showed model-specific benefits. For most primary QA models in both the DTA and Intervention settings, MAD improved over the QA baselines on most performance metrics except `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`'s performances are more mixed. It recorded a slight drop in MAP and WSS, but performed slightly better on recall. Comparatively, `Gemini-MAD`'s performance improvement 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. This results in two insights that may be useful for designing MAD systems. One the one hand, *multi-agent debate brings performance gains*, even with agents that are much weaker. On the other hand, *benefit brought by MAD also seem to be limited by other agents' capabilities* and weaker agents will likely bring less benefits to stronger ones.

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 0.1818 and 0.4067 to 0.2856 and 0.5403, respectively, while on Intervention from 0.2903 and 0.4296 to 0.4191 and 0.5994. It is worth noting that `Haiku-MAD` recorded the second performances on Intervention across all metrics and was only defeated by `Soft-Vote` by a tiny margin. The results suggest that *weaker agents may benefit more from exposure to alternative perspectives*, although it is unclear whether it is the nature of the different LLM architectures that causes this behavioral difference. In reverse, *opinions from stronger peers may have more significant impacts* on self-reflection and decision-making. Both conform well with common sense. Finally, the debating models, except `GPT-MAD`, exhibit good performances almost on par with the much more expensive adjudication methods, which highlights a promising future of MAD systems in systematic review automation.

Between LLM-based adjudication approaches, the adjudicator-as-a-ranker method `Adj-Rank` outperformed the adjudicator-as-a-judge approach `Adj-Judge` in most important metrics including MAP, R@50% and WSS@95%, demonstrating that *preserving diverse perspectives through weighted averaging* is 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 settingsexcept that both adjudicator variants were outperformed by `Haiku-MAD`. These results nonly only underscore the potential value of the increasingly popular "LLM-as-a-Judge" paradigm [29] and the superiority of adjudicator-as-a-ranker as a novel contribution to this paradigm, but also highlight the high potential of the multi-agent debating paradigm, especially when cost-effectiveness is considered. Note that our approach can be seen as an example of idea of exchange-of-thought [57]

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 approachwhere we see that any two-model combination also achieved substantial improvements over individual models. Particularly, GPT+Haiku performed surprisingly well (MAP: 0.46, WSS@95%: 0.57, WSS@100%: 0.49), and almost rivaled `Soft-Vote` on Intervention in some metrics (MAP: 0.46, WSS@95%: 0.61, WSS@100%: 0.53). These findings suggest that benefits from

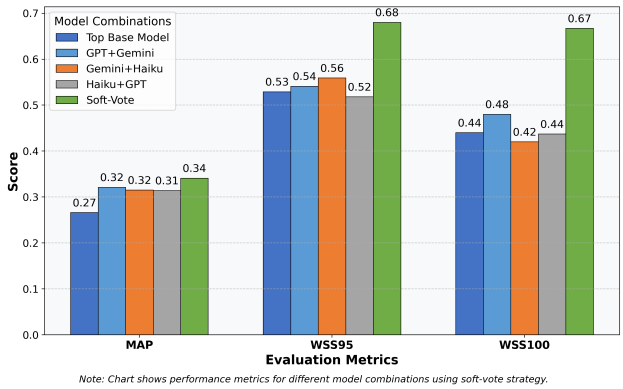
Table 1. Performances of Multi-LLM Collaboration

Setting	Models	MAP	R@5%	R@10%	R@50%	WSS@95%	WSS@100%
DTA	GPT	0.2714 (±0.1773)	41.74% (±0.2983)	56.41% (±0.3073)	89.65% (±0.1492)	0.4281 (±0.2586)	0.3617 (±0.3075)
	Gemini	0.2656 (±0.1979)	47.86% (±0.3252)	60.76% (±0.2945)	92.24% (±0.1217)	0.5290 (±0.3112)	0.4404 (±0.3700)
	Haiku	0.1818 (±0.1514)	33.21% (±0.3030)	46.99% (±0.2605)	90.15% (±0.1332)	0.4067 (±0.1858)	0.2876 (±0.2348)
	Soft-Vote	**0.3409 (±0.1737)**	**46.92% (±0.3105)**	**64.67% (±0.2968)**	**94.84% (±0.0924)**	**0.6798 (±0.2389)**	**0.6668 (±0.2786)**
	GPT-MAD	0.2705 (±0.1728)	40.47% (±0.2889)	59.78% (±0.3100)	90.22% (±0.1427)	0.5344 (±0.3201)	0.3775 (±0.3839)
	Gemini-MAD	0.2762 (±0.1927)	44.99% (±0.2944)	61.80% (±0.2919)	92.08% (±0.1355)	0.5727 (±0.2850)	0.4450 (±0.3878)
	Haiku-MAD	0.2856 (±0.1856)	47.91% (±0.3540)	61.27% (±0.3180)	91.14% (±0.1445)	0.5403 (±0.3143)	0.4495 (±0.3796)
	MAD-Soft-Vote	0.3279 (±0.2219)	49.11% (±0.4161)	64.66% (±0.3547)	91.65% (±0.1508)	0.5794 (±0.3490)	0.4556 (±0.4446)
	Adj-Judge	0.2665 (±0.1452)	48.63% (±0.2546)	66.13% (±0.2277)	91.71% (±0.1011)	0.5475 (±0.2352)	0.4341 (±0.2850)
	Adj-Rank	0.3445 (±0.2176)	49.79% (±0.3898)	64.22% (±0.3793)	93.57% (±0.1271)	0.5927 (±0.3452)	0.5001 (±0.4253)
	Adj-Soft-Vote	0.3516 (±0.2201)	49.43% (±0.3918)	64.63% (±0.3703)	93.57% (±0.1271)	0.5937 (±0.3453)	0.5151 (±0.4116)
Intervention	GPT	0.3954 (±0.1696)	46.01% (±0.1702)	62.32% (±0.1588)	90.87% (±0.0876)	0.4636 (±0.1988)	0.3920 (±0.2104)
	Gemini	0.3891 (±0.1587)	48.57% (±0.1846)	62.49% (±0.1590)	93.57% (±0.0599)	0.4808 (±0.1581)	0.4131 (±0.1892)
	Haiku	0.2903 (±0.1543)	35.49% (±0.1443)	52.64% (±0.1559)	93.74% (±0.0408)	0.4296 (±0.1518)	0.3440 (±0.1837)
	Soft-Vote	**0.4621 (±0.1673)**	**53.15% (±0.1875)**	**66.71% (±0.1649)**	**96.82% (±0.0360)**	**0.6064 (±0.1400)**	**0.5274 (±0.1727)**
	GPT-MAD	0.3760 (±0.1707)	48.50% (±0.1629)	62.06% (±0.1573)	92.30% (±0.0608)	0.4490 (±0.1933)	0.3885 (±0.2042)
	Gemini-MAD	0.4017 (±0.1891)	49.87% (±0.1727)	63.37% (±0.1433)	93.23% (±0.0613)	0.5087 (±0.1779)	0.4389 (±0.1972)
	Haiku-MAD	0.4191 (±0.1681)	54.43% (±0.1921)	68.01% (±0.1582)	96.02% (±0.0459)	0.5994 (±0.1677)	0.5363 (±0.2075)
	MAD-Soft-Vote	0.4563 (±0.1739)	53.44% (±0.1815)	68.49% (±0.1661)	96.28% (±0.0473)	0.5890 (±0.1611)	0.5269 (±0.1994)
	Adj-Judge	0.4273 (±0.1721)	53.17% (±0.1837)	67.29% (±0.1542)	93.19% (±0.0555)	0.5166 (±0.1839)	0.4764 (±0.1993)
	Adj-Rank	0.4521 (±0.1583)	50.50% (±0.1819)	65.57% (±0.1518)	94.32% (±0.0525)	0.5253 (±0.1688)	0.4623 (±0.2209)
	Adj-Soft-Vote	0.4627 (±0.1653)	51.63% (±0.1890)	65.85% (±0.1574)	95.68% (±0.0426)	0.5894 (±0.1595)	0.5308 (±0.1923)

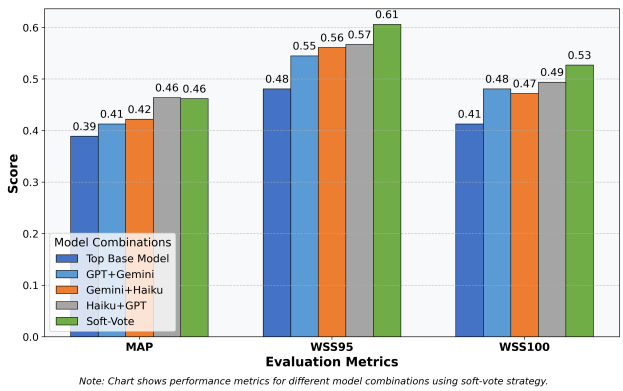
All values are reported as mean (± standard deviation).

Table 2. LLM Pricing and Cost Breakdown.

Setting	Type	Model	Input Price	Input Cost	Output Price	Output Cost	Total Cost
DTA	Question Answering	GPT-4o Mini	0.15	$2.54	0.6	$8.54	$11.08
		Gemini 1.5 Flash	0.075	$1.27	0.3	$4.17	$5.44
		Claude 3 Haiku	0.25	$4.24	1.25	$23.66	$27.90
	Majority Voting	`Soft-Vote`	-	$8.05	-	$36.38	$44.43
	Debating	GPT-4o Mini	0.15	$18.32	0.6	$47.38	$65.70
		Gemini 1.5 Flash	0.075	$13.19	0.3	$42.64	$55.82
		Claude 3 Haiku	0.25	$25.16	1.25	$65.15	$90.31
	Adjudication	Gemini 1.5 Pro	3.5	$263.49	10.5	$389.65	$653.15
Intervention	Question Answering	GPT-4o Mini	0.15	$4.25	0.6	$12.56	$16.81
		Gemini 1.5 Flash	0.075	$2.12	0.3	$5.55	$7.67
		Claude 3 Haiku	0.25	$7.08	1.25	$34.65	$41.73
	Majority Voting	`Soft-Vote`	-	$13.45	-	$52.76	$66.21
	Debating	GPT-4o Mini	0.15	$28.77	0.6	$68.74	$97.51
		Gemini 1.5 Flash	0.075	$21.11	0.3	$61.76	$82.88
		Claude 3 Haiku	0.25	$38.98	1.25	$94.45	$133.43
	Adjudication	Gemini 1.5 Pro	3.5	$394.11	10.5	$565.80	$959.90



(a) DTA setting



(b) Intervention setting

Fig. 1. Ablation Study of Majority Voting.

diversity begin with just two base models and increase as more models are added.

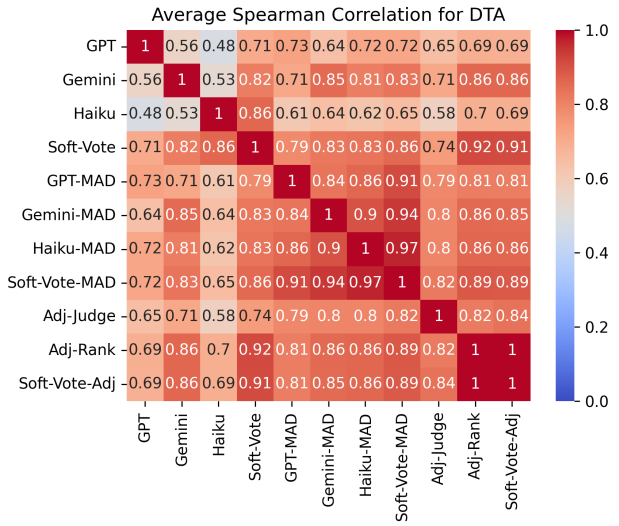
Correlation analysis in Figure 2 reveals moderate correlations between the QA models (Spearman's Rank Correlations: 0.48–0.56 on DTA and 0.49–0.52 on Intervention), indicating each primary model captures different aspects of relevance—a diversity that an ensemble approach can effectively leverage. `Soft-Vote` 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 Sect. 4.2.4.

The strong performance combined with computational efficiency (see the breakdown in Table 2), which is less than 1/14 of adjudication and at msot around 1/186 of the cost of a single human reviewer (detailed in the Supplementary Materials). 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.

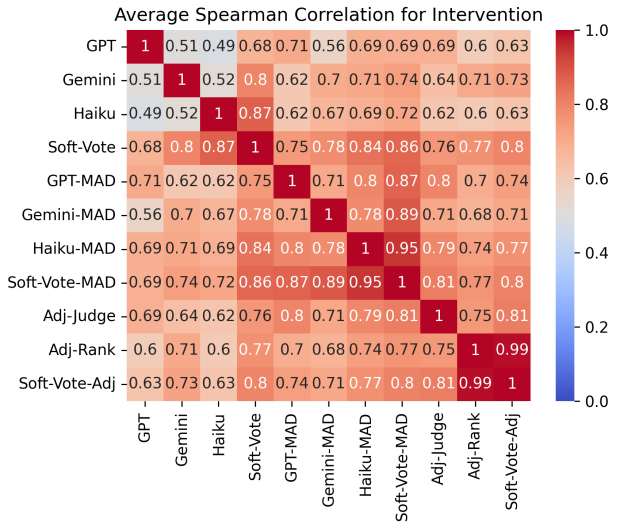
Critics of Multi-Agent Debate

The debating strategy aims to improve screening performance by leveraging collective intelligence through exchanging structured argument. Our analysis reveals model-dependent outcomes as shown in Figure 3. `Gemini-MAD` showed minimal to negligible improvement over the QA baseline, possibly due to its high baseline performance, less influence from external arguments, or suboptimal prompt tailoring. `GPT-MAD` exhibited more varied results. Despite of obvious improvements on DTA, `GPT-MAD` stumbled at improving screening performance across various metrics. Comparatively, `Haiku-MAD` obtained pronounce gains from debating. Although Haiku is notably the weakest QA model in medical abstract screening and much weaker than the other two competitors, it demonstrates a clear capacity for reasoning refinement through interaction. This phenomenon may have shed some *unusual light on the design of cost-effective LLM-MAD systems* because stronger individual models may not always benefit most from multi-agent debate.

The success genes of `Haiku-MAD` may to some extent lie in `Haiku` being more "diverse" from `GPT` (seen from the its comparatively low correlations with the latter: for example 0.48 and 0.53 on DTA in Figure 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



(a) DTA setting



(b) Intervention setting

Fig. 2. Correlation Analysis Between Models.

perhaps takes more peer opinions than its own, potentially implied from the fact that `Haiku-MAD` is more correlated with `GPT` and `Gemini` rather than with `Haiku`. The findings may have an important implication: *It might not be the individual agent's own capability but its capability of assimilating different opinions that makes MAD systems.*

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 debating models, which further demonstrates the *core role of model heterogeneity* [58].

Table 3. Ratings of QA Models by the Adjudicator.

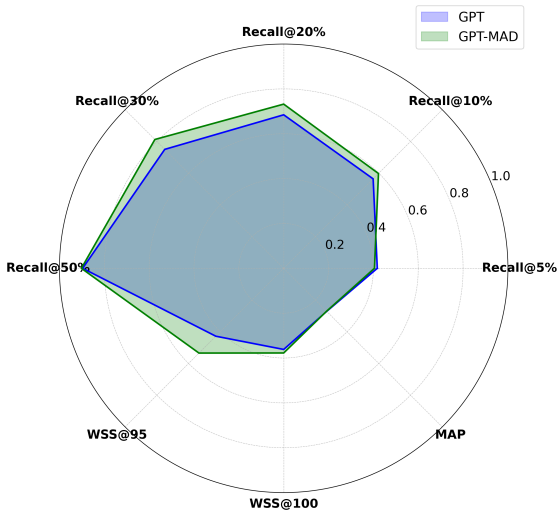
Setting	Model	Avg Rating	Best (%)	Worst (%)
DTA	`GPT`	0.855 (±0.018)	9.86%	21.62%
	`Gemini`	0.917 (±0.019)	51.92%	10.17%
	`Haiku`	0.762(±0.021)	14.40%	28.96%
Intervention	`GPT`	0.887 (±0.016)	14.45%	12.98%
	`Gemini`	0.885 (±0.031)	45.15%	19.33%
	`Haiku`	0.760 (±0.086)	10.71%	34.09%

Critics of Adjudication

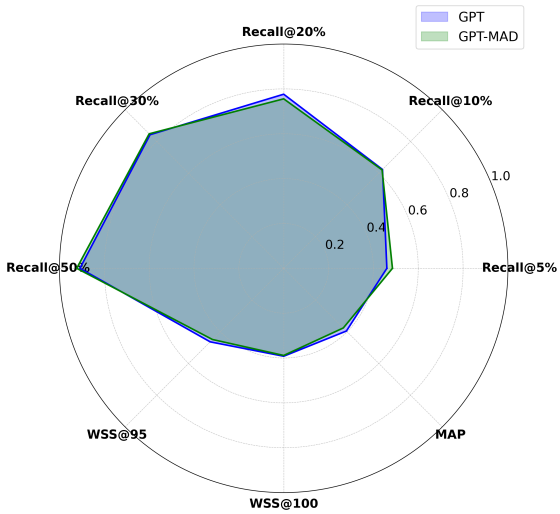
Adjudication strategies introduce a hierarchical decision-making layer in a multi-LLM collaborative framework. Table 3 shows the average score assigned to each primary QA model by the Adjudicator model (here Gemini 1.5 Pro), along with how often each QA model was deemed as the best or worst performing model by the adjudicator. Sect. 4.1 shows `Gemini` (Gemini 1.5 Flash) as 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", which makes sense because it is indeed the weakest screening 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 likely because `Gemini`'s performance gain over other models is less significant than on DTA. These findings have several implications. Firstly, the adjudicator may indeed have some good capabilities in deciding the more appropriate primary model on a case by case basis, a key for significantly improved screening performances by both adjudication methods. In the meanwhile, they have also revealed *a potential bias of the adjudicator towards its own model family*, in corroboration with findings reported in [38]. This is also reflected by the often stronger correlations between `Gemini` and the adjudicators (Figure 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 preferences or limit the diversity of perspectives.

Model Diversity

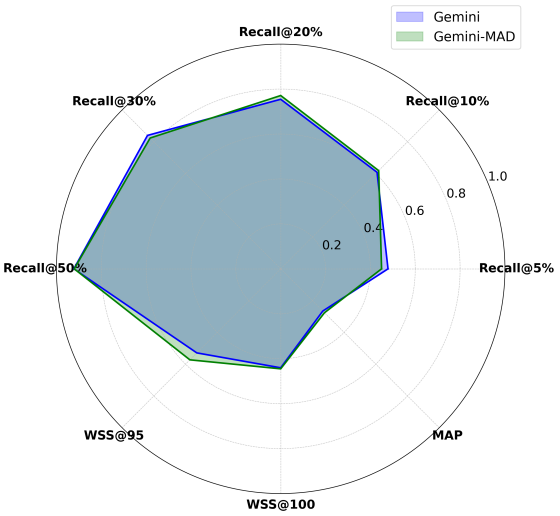
Model diversity plays a central role in ensemble performance. The `Soft-Vote` approach, combining three distinct QA models, benefits from a "healthy" inter-model disagreement (i.e., model heterogenity [58]), enabling it to balance strengths and mitigate weaknesses across agents. `MAD-Soft-Vote` in Table 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-MAD`. This leads to marginal gains of `MAD-Soft-Vote` over 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. In summary, while both debate and adjudication offer refinements, preserving



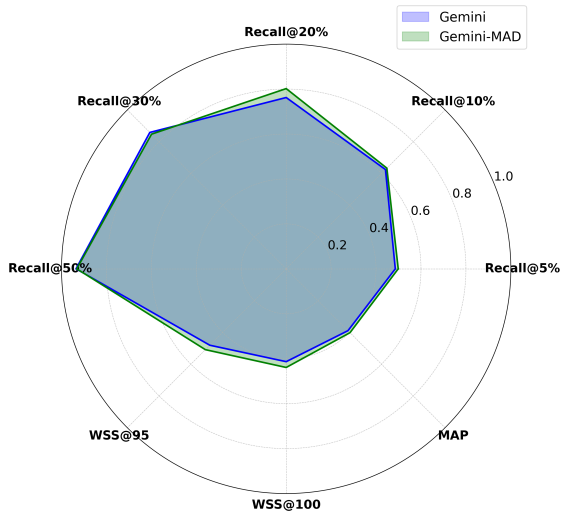
(a) GPT-MAD vs. GPT on DTA



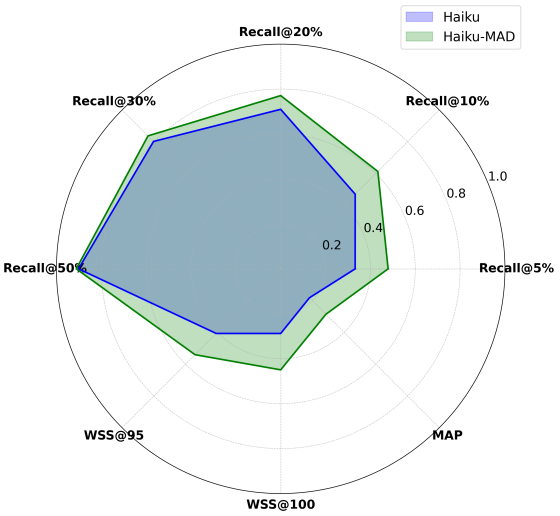
(b) GPT-MAD vs. GPT on Intervention



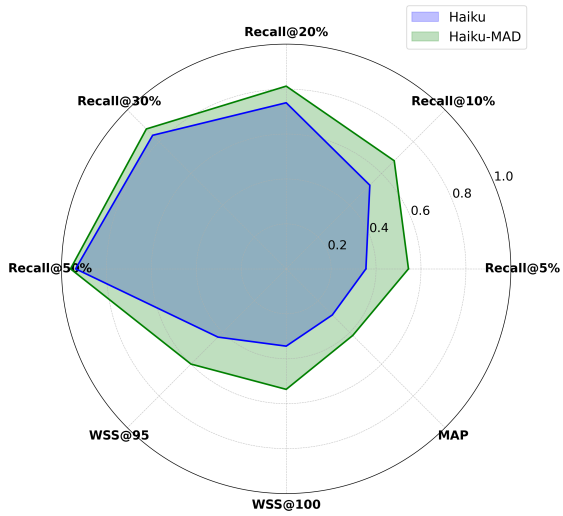
(c) Gemini-MAD vs. Gemini on DTA



(d) Gemini-MAD vs. Gemini on Intervention



(e) Haiku-MAD vs. Haiku on DTA



(f) Haiku-MAD vs. Haiku on Intervention

Fig. 3. Multi-Agent Debate vs. QA Models

model diversity through simple ensembling (`Soft-Vote`) remains the most cost-effective and robust strategy under resource constraints.

Re-Ranking Results

Table 4 shows the improved performances using both macro- and micro-level re-ranking (for simplicity of comparison, $\alpha = \beta = 0.5$), including two important baselines from [2]. `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 [2] that integrates both macro- and micro-level re-ranking, thus comparable to our re-ranking variants.

`Soft-Vote w/o re-ranking` significantly outperformed the best results in [2], particularly in WSS. Although GPT-3.5 was used in [2], 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` significantly improved over `Soft-Vote w/o re-ranking` on both DTA and Intervention. Notably, re-ranking also consistently improved the performances of other collaborative strategies by large margins, almost rivaling `Soft-Vote w/ re-ranking` on DTA.

Conclusion

This study presents an in-depth investigation into multi-LLM collaborative strategies for automating abstract screening. We successfully developed and evaluated three collaborative strategies–majority voting, multi-agent debate and LLM-based adjudication, and demonstrated that collaboration among multiple, cost-effective LLMs has high potential to substantially reduce screening workload and costs.

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 or heterogeneity on the success of effectively aggregating relatively weaker screening models. While debating improves screening performance, the 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 multi-agent 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 1/16 of that of adjudication methods based on a strong LLM, and achieves more than 185× cost reduction compared to single human reviewer based on the most conservative estimation according to British academic salary scales. By demonstrating that multi-LLM collaboration can achieve superior performance at substantially lower costs, this research offers a pathway toward making systematic review automation both more effective and affordable. The collaborative frameworks effectively mitigate individual LLMs' biases through collective intelligence.

Table 4. Performance Improvements by Re-ranking.

Setting	Model	MAP	R@5%	R@10%	R@20%	R@30%	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	0.3409 (±0.1737)	46.92% (±0.3105)	64.67% (±0.2968)	79.31% (±0.2686)	86.27% (±0.2317)	94.84% (±0.0924)	0.6798 (±0.2389)	0.6668 (±0.2786)
	Soft-Vote w/ re-ranking	**0.3595 (±0.1459)**	**56.52% (±0.3609)**	**72.32% (±0.3233)**	**84.67% (±0.2602)**	**88.59% (±0.2440)**	**96.59% (±0.0845)**	**0.7083 (±0.2299)**	**0.6635 (±0.2716)**
	GPT-MAD w/ re-ranking	0.3478 (±0.1482)	57.47% (±0.3748)	71.36% (±0.3208)	82.86% (±0.2714)	89.71% (±0.2195)	96.91% (±0.0854)	0.6904 (±0.2461)	0.6483 (±0.2726)
	Gemini-MAD w/ re-ranking	0.3758 (±0.1451)	55.84% (±0.3774)	71.63% (±0.3232)	83.17% (±0.2817)	88.70% (±0.2305)	96.62% (±0.0940)	0.7039 (±0.2260)	0.6549 (±0.2633)
	Haiku-MAD w/ re-ranking	0.3681 (±0.1539)	57.66% (±0.3671)	71.62% (±0.3090)	82.67% (±0.2774)	87.76% (±0.2459)	96.33% (±0.1026)	0.7049 (±0.2301)	0.6541 (±0.2558)
	Adj-Judge w/ re-ranking	0.3635 (±0.1405)	57.40% (±0.3649)	71.92% (±0.3291)	82.93% (±0.2793)	89.17% (±0.2274)	97.09% (±0.0860)	0.6968 (±0.2309)	0.6737 (±0.2507)
	Adj-Rank w/ re-ranking	0.3782 (±0.1408)	56.85% (±0.3829)	71.41% (±0.3333)	84.11% (±0.2663)	88.62% (±0.2460)	96.59% (±0.0845)	0.7062 (±0.2361)	0.6669 (±0.2678)
Intervention	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	0.4621 (±0.1673)	53.15% (±0.1875)	66.71% (±0.1649)	83.41% (±0.1007)	90.11% (±0.0688)	96.82% (±0.0360)	0.6064 (±0.1400)	0.5274 (±0.1727)
	Soft-Vote w/ re-ranking	**0.4703 (±0.1586)**	**56.59% (±0.2065)**	**71.76% (±0.1674)**	**84.89% (±0.1109)**	**91.48% (±0.0802)**	**97.87% (±0.0264)**	**0.6961 (±0.1342)**	**0.6356 (±0.1857)**
	GPT-MAD w/ re-ranking	0.4469 (±0.1627)	56.11% (±0.2046)	70.06% (±0.1778)	83.16% (±0.1294)	89.96% (±0.0786)	97.48% (±0.0356)	0.6580 (±0.1441)	0.6092 (±0.1838)
	Gemini-MAD w/ re-ranking	0.4696 (±0.1717)	55.52% (±0.2082)	71.63% (±0.1682)	84.52% (±0.1201)	91.40% (±0.0802)	97.39% (±0.0364)	0.6726 (±0.1443)	0.6331 (±0.1780)
	Haiku-MAD w/ re-ranking	0.4585 (±0.1583)	55.78% (±0.2021)	70.17% (±0.1767)	84.53% (±0.1265)	91.46% (±0.0830)	98.01% (±0.0271)	0.6896 (±0.1386)	0.6427 (±0.1865)
	Adj-Judge w/ re-ranking	0.4821 (±0.1589)	58.08% (±0.2097)	72.05% (±0.1685)	84.55% (±0.1153)	90.42% (±0.0764)	97.57% (±0.0340)	0.6629 (±0.1421)	0.6155 (±0.1856)
	Adj-Rank w/ re-ranking	0.4461 (±0.1653)	54.11% (±0.1951)	69.32% (±0.1726)	82.96% (±0.1225)	89.92% (±0.0786)	96.97% (±0.0399)	0.6558 (±0.1448)	0.5851 (±0.1830)

References

1. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
2. Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. A question-answering framework for automated abstract screening using large language models. *Journal of the American Medical Informatics Association*, 31(9):1939–1952, 2024.
3. Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
4. Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545, 2017.
5. Wichor M Bramer, Melissa L Rethlefsen, Jos Kleijnen, and Oscar H Franco. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic reviews*, 6:1–12, 2017.
6. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
7. Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, Matthew Cecere, Anabel Selemon, Jason Sang, Ling Xi Gong, Robert Kloosterman, Scott Jiang, Richard Saleh, Denis Margalik, James Lin, Jane Jomy, Jerry Xie, David Chen, Jaswanth Gorla, Sylvia Lee, Kelvin Zhang, Harriet Ware, Mairead Whelan, Bijan Teja, Alexander A. Leung, Lina Ghosn, Rahul K. Arora, Allen S. Detsky, Michael Noetel, David B. Emerson, Isabelle Boutron, David Moher, George Church, and Niklas Bobrovitz. Automation of systematic reviews with large language models. *medRxiv*, 2025.
8. Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
9. Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
10. Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.
11. Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.
12. Aaron M Cohen, Kyle Ambert, and Marian McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association*, 16(5):690–704, 2009.
13. Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 2006.
14. Gordon V Cormack and Maura R Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868*, 2015.
15. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
16. Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
17. Phil Edwards, Mike Clarke, Carolyn DiGuiseppi, Sarah Pratap, Ian Roberts, and Reinhard Wentz. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in medicine*, 21(11):1635–1640, 2002.
18. Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.
19. Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.
20. Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Mike Paget, and Christopher Naugler. Automated paper screening for clinical reviews using large language models. *arXiv preprint arXiv:2305.00844*, 2023.
21. Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.
22. Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Survey Track.
23. Hannah Harrison, Simon J Griffin, Isla Kuhn, and Juliet A Usher-Smith. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC medical research methodology*, 20:1–12, 2020.
24. Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
25. Mark Hughes, Irene Li, Spyros Kotoulas, and Toyotaro Suzumura. Medical text classification using convolutional neural networks. In *Informatics for health: connected citizen-led wellness and population health*, pages 246–250.

IOS Press, 2017.
26. Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.
27. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.
28. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
29. Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025.
30. Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.
31. Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024.
32. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
33. Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models, 2024.
34. Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
35. Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253, 2014.
36. Hanna Olofsson, Agneta Brolund, Christel Hellberg, Rebecca Silverstein, Karin Stenström, Marie Österberg, and Jessica Dagerhamn. Can abstract screening workload be reduced using text mining? user experiences of the tool rayyan. *Research synthesis methods*, 8(3):275–280, 2017.
37. Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1–22, 2015.
38. Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc., 2024.
39. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
40. John Rathbone, Matt Carter, Tammy Hoffmann, and Paul Glasziou. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. *Systematic reviews*, 4:1–6, 2015.
41. Rohan Sanghera, Arun James Thirunavukarasu, Marc El Khoury, Jessica O'Logbon, Yuqing Chen, Archie Watt, Mustafa Mahmood, Hamid Butt, George Nishimura, and Andrew A S Soltan. High-performance automated abstract screening with large language model ensembles. *Journal of the American Medical Informatics Association*, 32(5):893–904, 03 2025.
42. Burr Settles. Active learning literature survey. 2009.
43. Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc., 2023.
44. Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
45. James Thomas, John McNaught, and Sophia Ananiadou. Applications of text mining within systematic reviews. *Research synthesis methods*, 2(1):1–14, 2011.
46. Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. Systematic review automation technologies. *Systematic reviews*, 3:1–15, 2014.
47. Sai Krishna Vallamchetla, Omar Abdelkader, Ali Elnaggar, Doaa Ramadan, Md Manjurul Islam Shourav, Irbaz B. Riaz, and Michelle P. Lin. Do it faster with picos: Generative ai-assisted systematic review screening. *Journal of Biomedical Informatics*, 168:104860, 2025.
48. Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133, 2021.
49. Stevie Van der Mierden, Katya Tsaioun, André Bleich, and Cathalijn HC Leenaars. Software tools for literature screening in systematic reviews in biomedical research. *Altex*, 36(3):508–517, 2019.
50. Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182, 2010.
51. Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. Semi-automated

screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):1–11, 2010.

52. Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
53. Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 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*, pages 1426–1436, 2023.
54. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
55. Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.
56. Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.
57. Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore, December 2023. Association for Computational Linguistics.
58. Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity, 2025.
59. Zhihong Zhang, Mohamad Javad Momeni Nezhad, Pallavi Gupta, Ali Zolnour, Hossein Azadmaleki, Maxim Topaz, and Maryam Zolnoori. Enhancing ai for citation screening in literature reviews: Improving accuracy with ensemble models. *International Journal of Medical Informatics*, 203:106035, 2025.