



Collaborating with large language models in literature screening for a systematic review of college students' GenAI literacy

Wonchan Choi, Joyce Lee, Besiki Stvilia, Yan Zhang, and Hyerin Bak

DOI: <https://doi.org/10.47989/ir31iConf64265>

Abstract

Introduction. Literature screening is among the most time- and labour-intensive phases of systematic literature reviews (SLRs). Although large language models (LLMs) have been explored for screening, prior work has focused on medical and environmental domains and mainly benchmarked LLMs against human coders, offering limited guidance on collaborative integration in SLRs.

Method. This study evaluated 12 GPT model-prompt configurations in an SLR of 1,616 publications. Two human coders screened a 10% sample ($n = 162$) to create a gold standard for model comparison. Performance was assessed using balanced accuracy, recall for inclusion, time efficiency, and cost efficiency. Disagreements between humans and models were analysed.

Analysis. Bootstrap tests compared performance across configurations. Open coding identified error types in human-model disagreements. Group discussion resolved discrepancies.

Results. GPT-5 Zeroshot and GPT-4o-mini Fewshot achieved the highest performance (accuracy = 0.990 and 0.946; recall = 1.000 and 0.933). GPT-4o-mini was faster and cheaper but more prone to overly rigid rule application. Error analysis identified 10 mismatches, leading to two corrections of human miscoding.

Conclusion. LLM-assisted screening can reduce workload, improve efficiency, and correct human errors in SLR. Practical guidelines for prompt design and confidence thresholds can position LLMs as collaborative tools in SLRs.

Introduction

Emerging artificial intelligence (AI) tools have great potential in automating repetitive tasks performed by people or that are error prone. The literature screening process for systematic literature reviews (SLRs), in which researchers manually select literature based on inclusion and exclusion criteria, is notably time- and labor-intensive (Guo et al., 2024; Issaiy et al., 2024). Few studies outside medical and environmental disciplines have used large language models (LLMs) for SLRs; also, few ongoing SLRs have used LLMs to date. Several experiments were conducted after SLRs, in which researchers compared LLM and human screening results, with the latter as a gold standard (Matsui et al., 2024). In these studies, the LLM models were not limited to OpenAI's ChatGPT models; Google's Bard, Meta's Llama 2, Anthropic's Claude AI, and other semi-automated AI tools like ASReview, Rayyan, or Abstrackr (Ghossein et al., 2025; Guo et al., 2024; Matsui et al., 2024; Syriani et al., 2024; Van Dijk et al., 2023) have been used for SLRs.

Studies using ChatGPT (e.g., GPT-3, GPT-4 Turbo) have shown that well-structured prompts can reduce screening time, though concerns remain about reproducibility, especially for interdisciplinary topics (Ghossein et al., 2025; Nykvist et al., 2025; Oami et al., 2024). Prompt-engineering strategies, such as chain-of-thought reasoning (Issaiy et al., 2024), layered prompting (Matsui et al., 2024), and multiple runs of refined prompts (Syriani et al., 2024), improved consistency, but models often struggled with false negatives—erroneous exclusions of relevant studies (Matsui et al., 2024; Van Dijk et al., 2023). Beyond prompt refinement, fine-tuning of model configurations has been explored. This includes adjusting hyperparameters, such as setting temperature to 0 for deterministic output (Chibwe et al., 2025) and addressing stochastic variability through repeated runs and majority voting across outputs to improve reliability (Syriani et al., 2024).

SLRs in fast-moving areas face heavy screening loads. Recall must stay high, yet criteria are subtle and easy to misapply. Although previous studies have explored innovative ways of using LLMs for literature screening, practical guidelines for integrating them into SLRs as collaborative partners remain underexplored. More work is needed to show how LLM outputs can be governed, audited, and fed into the analysis protocol. As part of a larger project on students' generative AI (GenAI) literacy in higher education, this study addressed this gap by synthesising and applying established LLM-based screening practices to an interdisciplinary SLR context and evaluating their use through two research questions (RQs):

1. How do different LLMs compare in terms of classification accuracy, time efficiency, and cost efficiency?
2. What are the nature and sources of disagreements between LLMs and humans?

Method

Based on best practices reported in the literature (Guo et al., 2024; Matsui et al., 2024; Syriani et al., 2024; Zuo et al., 2025), this study developed and evaluated a workflow for applying LLMs to title and abstract screening in a SLR on college students' GenAI literacy. The procedure consisted of five phases: (1) literature search and deduplication; (2) dataset construction and gold standard development; (3) GPT model configuration; (4) GPT model evaluation; and (5) discrepancy and error pattern analysis (Figure 1).

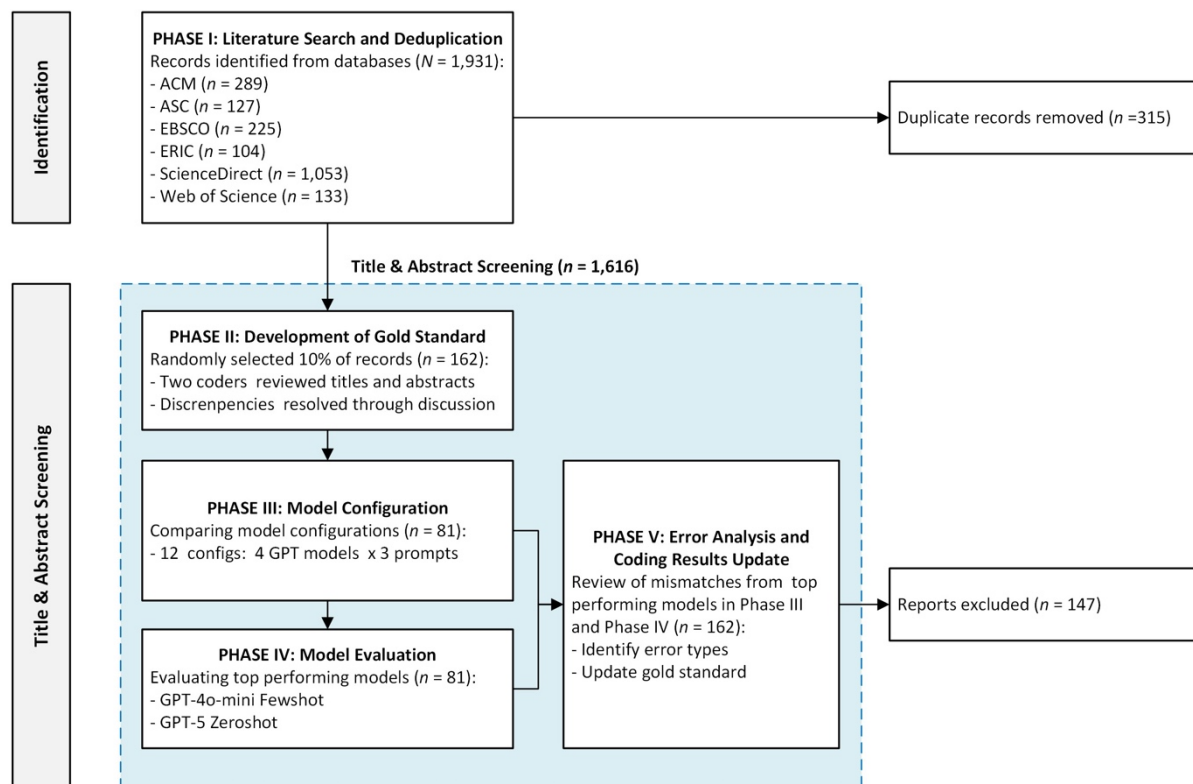


Figure 1. Workflow.

Phase 1: literature search and deduplication

We searched five electronic databases: ACM Digital Library, EBSCO Academic Search Complete, EBSCO ERIC, ScienceDirect, and Web of Science. The Boolean query combined three keyword groups: (a) GenAI terms ('generative artificial intelligence' OR 'GenAI' OR 'generative AI' OR 'GAI'); (b) population terms ('college students' OR 'university students' OR 'higher education'); and (c) competence terms: (literac* OR competenc*). Queries were adapted to each interface.

Only original research articles in journals and conference proceedings were included; reviews, editorials, and abstracts were excluded. Because ChatGPT's first release was in 2022, publication years were limited to 2022–2025. The final search (August 5, 2025) returned 1,931 records. After removing duplicates in Rayyan, a reliable tool for SLR (Guimarães et al., 2022), 1,616 unique articles were retained for screening.

Phase 2: development of gold standard

To provide a reliable basis for evaluating LLM performance, we created a gold standard dataset by randomly sampling 162 articles, representing approximately 10% of the corpus. This proportion is consistent with the range commonly used to develop and test automated screening methods (Zuo et al., 2025). Two researchers independently screened these articles using predefined eligibility rules. Articles were included if they: (a) addressed college students' needs, skills, or literacy related to GenAI; (b) focused on higher education contexts; and (c) reported original empirical data collected directly from U.S. college students. Articles were excluded if they: (a) were not original research (e.g., reviews, editorials, or commentaries); (b) relied solely on secondary or simulated data; (c) did not involve college students as the primary population; (d) examined technologies other than GenAI; or (e) collected data outside the United States.

Intercoder reliability was high (Cohen's $\kappa = .87$, 95% CI [.72, .97]), indicating almost perfect agreement (Landis & Koch, 1977). The prevalence- and bias-adjusted kappa (.95, 95% CI [.90, .99])

confirmed reliability after adjusting for prevalence effects (Byrt et al., 1993). Disagreements were resolved through consensus to produce the gold standard benchmark for model evaluation. The 162-article set was divided equally into a configuration set ($n = 81$) for model configuration and an evaluation set ($n = 81$) for unbiased testing.

Phase 3: model configuration

This stage explored GPT configurations by varying factors highlighted in prior studies—model version and prompt engineering (Guo et al., 2024; Matsui et al., 2024; Zuo et al., 2025)—and identified the most suitable model based on three criteria: performance, cost, and efficiency. Python was used as the primary environment to run GPT models through OpenAI’s application programming interface and conduct analyses. The present paper reports on the results of 12 model configurations (combinations of four GPT models and three prompt-engineering strategies).

GPT models and parameter settings

We selected four GPT models based on their capacities and trade-offs: (a) **GPT-5**, the flagship reasoning model; (b) **GPT-5-mini**, a lightweight and cost-efficient counterpart of GPT-5; (c) **GPT-4.1**, the prior flagship model; and (d) **GPT-4o-mini**, the model with the lowest computational cost. At the time of testing (August 2025), GPT-5 models did not permit parameter adjustments through OpenAI’s platform. For GPT-4 models, we fixed the temperature at 0 and set top_p to 0.4 based on previous studies to improve reproducibility and screening performance (Matsui et al., 2024; Zuo et al., 2025).

Each model configuration was run three times to account for the stochastic variability of LLM outputs, and majority voting across runs was used to determine final model decisions (Chibwe et al., 2025; Matsui et al., 2024).

Prompt engineering

Guided by prior work (Matsui et al., 2024; Syriani et al., 2024), we developed a prompt that incorporated: (a) task description (inclusion or exclusion decision, justification, and confidence score [0–100]), (b) screening rules, (c) definitions of key terms (GenAI and higher education), and (d) inclusion and exclusion criteria (Figure 2). To mitigate false negatives, the prompt explicitly instructed models to treat uncertain cases as potentially eligible rather than excluded, aligning with prior recommendations (Nykivist et al., 2025; Syriani et al., 2024).

You are a meticulous researcher screening titles and abstracts of peer-reviewed articles for a systematic review.

Your task:
 Classify the article as Included or Excluded.
 Return exactly three lines without punctuation labels:
 [Your decision]
 [Reason – A brief, specific explanation referencing the inclusion and exclusion criteria below]
 [Confidence 0-100 as an integer]

Screening Rules:
 Treat missing or ambiguous information as uncertainty that should NOT count against inclusion. When uncertain, prefer including rather than excluding.
 Exclude only if the title or abstract clearly meets one or more exclusion criteria.

Definitions:
 Generative AI (GenAI) refers to large-scale models that generate novel text, code, images, audio, or video (e.g., ChatGPT, Gemini, GitHub Copilot, DALL-E). Tools limited to plagiarism detection, grading automation, rule-based chatbots, or template filling are NOT GenAI.
 Higher education refers to the stage of formal learning that occurs after completion of secondary education (high school), typically provided by colleges, universities, and other postsecondary institutions.

Inclusion Criteria
 1. Topic: The article addresses college students' needs, experiences, skills, competencies, or literacy related to GenAI.
 2. Context: The article focuses on the higher education.
 3. Geographic scope: The study reports original empirical data collected directly from college students in the United States. Studies with mixed participant groups are eligible only if findings for U.S. college students are reported separately.

Exclusion Criteria:
 1. Publication type is not original research (e.g., review, editorial, commentary, opinion, conceptual paper).
 2. Research relies on secondary data, simulations, or other methods without direct student input.
 3. Primary population is not college students (e.g., faculty, teachers, librarians, administrators, K-12 students).
 4. Study concentrates on tools outside the GenAI definition (e.g., plagiarism detection, grading automation, rule-based chatbots, or template filling, LMS platforms).
 5. Data were collected outside the United States. For mixed samples, exclude unless US data are reported separately.

Figure 2. Prompt.

Using this prompt as the foundation, we evaluated three prompt-engineering strategies: (a) **Zeroshot**, which provided only the prompt (Figure 2) without examples; (b) **Fewshot**, which supplemented the prompt with inclusion and exclusion cases; and (c) **Fewshot-Inc-Only**, which supplemented the prompt with only inclusion cases to minimise the risk of false negatives based on exclusion examples (Matsui et al., 2024). The primary goal of screening, especially in the initial title and abstract stage, is to maximise recall for inclusion.

Metrics

To assess model performance on the binary classification task (i.e., include vs. exclude), we employed balanced accuracy and recall for inclusion (hereafter, accuracy, and recall, respectively). Because the dataset was highly imbalanced, with excluded papers representing the majority—a common pattern in SLRs (Matsui et al., 2024; Syriani et al., 2024)—overall accuracy would overstate performance if a model simply rejected most studies. Balanced accuracy addresses this by averaging recall for both classes, capturing how well the model identifies included papers and rejects excluded ones.

Model selection

We used bootstrapping to estimate confidence intervals and conduct pairwise significance tests on differences in performance metrics—accuracy and recall. When configurations showed no significant differences across performance metrics, we prioritised the option with higher scores.

Phase 4: model evaluation

From the configuration results, we selected top-performing models and evaluated them against an independent dataset ($n = 81$). We used the same metrics (accuracy and recall) and applied pairwise bootstrap tests to assess whether performance differences were statistically significant.

Phase 5: discrepancy analysis

We analysed mismatches—in which model decisions diverged from the gold standard—across both the development and evaluation datasets ($N = 162$). To investigate sources of disagreement, we examined the models' justifications and confidence scores. Using open coding and group discussion, we identified recurring patterns. When models' reasoning aligned more closely with the inclusion or exclusion criteria than the human coders' judgments, we adopted the models' decisions and recalculated the performance metrics.

In addition to primary performance metrics, we incorporated efficiency measures—task completion time and cost per task (in USD)—as secondary criteria to assess model scalability.

Results

Model performance and efficiency (RQ1)

Figures 3 and 4 show accuracy and recall for the 12 model configurations. **GPT-4o-mini Fewshot** achieved the highest accuracy (.993), significantly outperforming its Zeroshot and Fewshot-Inc-Only, GPT-4.1 variants ($\Delta = .187-.194$, 95% CI [.042, .382], $p = .019-.026$), and GPT-5-mini Fewshot ($\Delta = .090$, 95% CI [.007, .229], $p = .014$). Its performance was statistically indistinguishable from **GPT-5 Zeroshot**, the other top-performing configuration ($\Delta = .000$, 95% CI [-.068, .068], $p = .500$).

For recall, **GPT-4o-mini Fewshot** again achieved the highest score (1.00), significantly outperforming GPT-4.1 Zeroshot and GPT-4o-mini Zeroshot ($\Delta = .375$, 95% CI [.125, .750], $p = .046$).

GPT-4o-mini Fewshot had numerically higher recall than **GPT-5 Zeroshot** ($\Delta = .125$, 95% CI [0.000, .375]), but the difference was not statistically significant ($p = .685$).

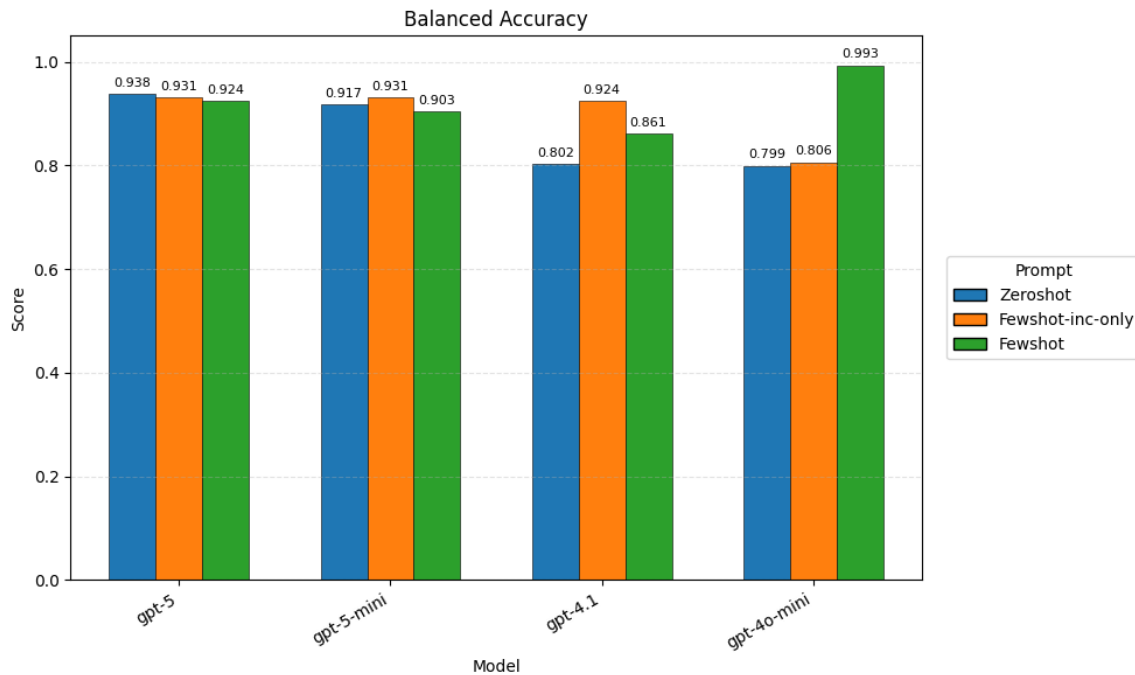


Figure 3. Balanced accuracy of all model configurations.

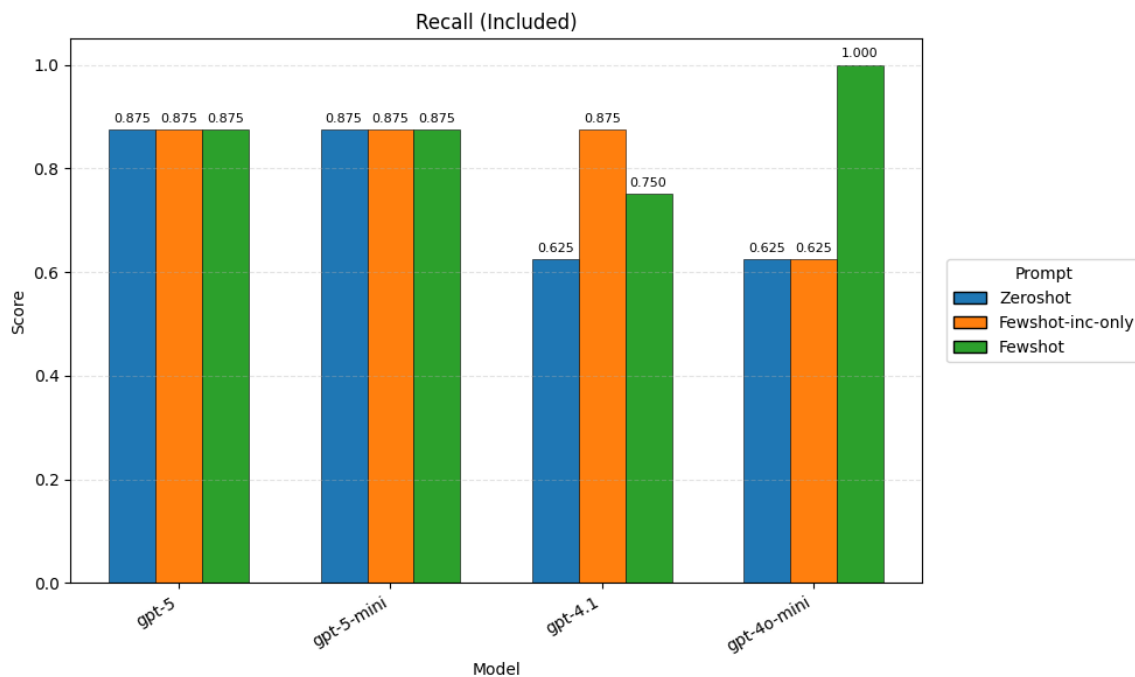


Figure 4. Recall for inclusion of all model configurations.

Performance of top two models

GPT-5 Zero-shot (accuracy = .973, 95% CI [.947, .994]) and **GPT-4o-mini Fewshot** (accuracy = .895, 95% CI [.717, .987]) performed at a similar level, with no statistically significant difference ($\Delta = .078$, 95% CI [-.235, .014], $p = .292$). For recall, **GPT-5 Zero-shot** achieved the maximum score of 1.00 (95% CI [1.000, 1.000]) compared to .857 (95% CI [.500, 1.000]) for **GPT-4o-mini Fewshot**, but the difference was not statistically significant ($\Delta = -.143$, 95% CI [-.429, .000], $p = .687$).

Cost and time efficiency

The two configurations differed substantially in efficiency. **GPT-5 Zeroshot** required about 54.8 minutes (20.3 seconds per article), whereas **GPT-4o-mini Fewshot** completed the task in 13.8 minutes (5.1 seconds per article). In terms of cost, **GPT-5 Zeroshot** charged \$2.00 for screening 162 articles, compared to \$0.12 for **GPT-4o-mini Fewshot**.

Disagreements and error types (RQ2)

Ten mismatches occurred between human coders and one or both top-performing GPT models across the two datasets (N = 162). In two cases, the research team determined that the models' decisions were correct, leading to updates in the gold standard.

The first case involved a study of Chinese international students, in which it was unclear whether participants had been recruited from U.S. colleges. Consistent with GPT-5's decision, the team concluded that the study should be retained for further review rather than excluded at the title and abstract stage, because references to '*Chinese international students*' or '*Chinese-background students*' did not provide sufficient information to apply the fifth exclusion criterion definitively (Figure 1). The second case concerned a study in which GenAI literacy appeared only in the discussion section. Although human coders had initially marked it for inclusion, the team agreed with GPT-5 that the study did not focus on GenAI use or literacy and should be excluded. After correcting these two human errors, nine errors remained, attributable to one or both models (Table 1).

Category	Error Type (ET)	GPT-5		GPT-4o-mini	
		n	Conf. Score	n	Conf. Score
False inclusion	[ET1] Confusion regarding scope of student roles: Both models wrongly included teaching-focused articles (e.g., training preservice teachers) despite the focus on students' learning-related GenAI literacy.	3	57.8	1	80.0
	[ET2] Overlook or misinterpretation of research design and evidence type: GPT-4o-mini failed to distinguish documentary or secondary sources (e.g., syllabi, policies, curricula) from empirical data collected directly from students (e.g., surveys, interviews, observations).	0		4	81.7
	[ET3] Overly rigid fixation on specific words in prompt: The model was influenced by peripheral wording rather than the central intent of each criterion. In the abstract, the phrase 'Future possibilities for students to seek feedback from generative artificial intelligence are briefly sketched' refers only to a speculative discussion on future directions, not empirical data. This should have led to exclusion, but the model included it.	0		1	78.3
False exclusion	[ET2] Overlook or misinterpretation of research design and evidence type; [ET3] Overly rigid fixation on specific words in the prompt: GPT-4o-mini overlooked that the study reported qualitative data from faculty members and students in higher education (focus groups and questionnaires). Also, although framed broadly as AI in higher education, the empirical focus of the study included students' perspectives, use cases, and challenges. Thus, the article should have been included, but the model excluded it.	0		1	85.0

Table 1. Error types in GPT model decisions.

With updates to the gold standard, both models showed improved metrics (Figure 5). **GPT-5 Zeroshot** (accuracy = .990, 95% CI [.977, 1.000]) and **GPT-4o-mini Fewshot** (accuracy = .946, 95% CI [.861, .990]) achieved high accuracy with no statistically significant difference ($\Delta = .044$, 95% CI [-.117, .003], $p = .119$). For recall, **GPT-5 Zeroshot** reached the maximum of 1.00 (95% CI [1.000, 1.000]), whereas **GPT-4o-mini** achieved .933 (95% CI [.769, 1.000]); this difference was nonsignificant ($\Delta = -.067$, 95% CI [-.200, .000], $p = .710$).

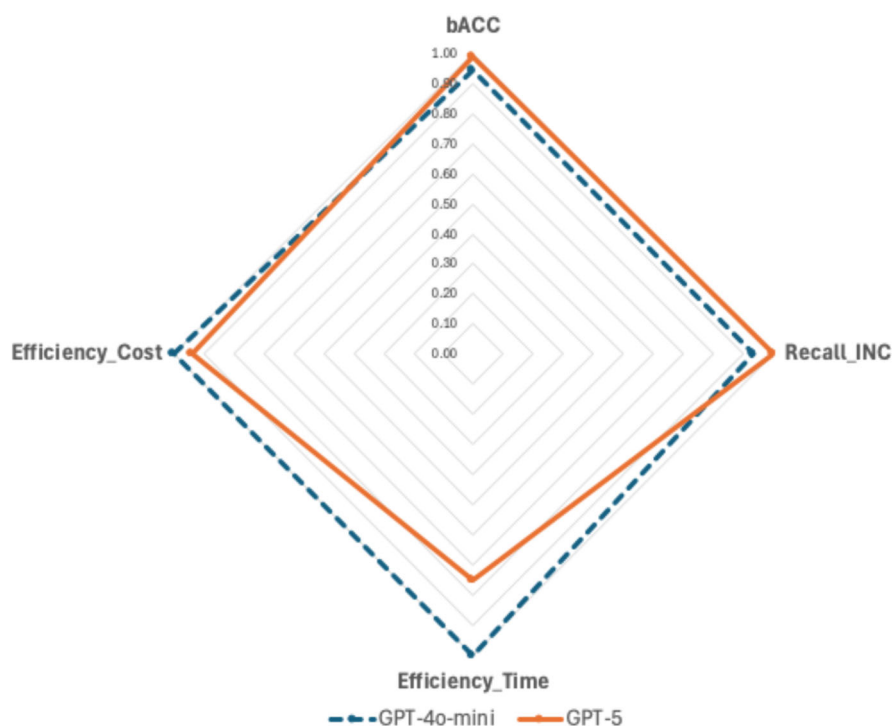


Figure 5. Revised metrics with normalised time and cost efficiency.

Discussion

Our findings show that both **GPT-5 Zeroshot** and **GPT-4o-mini Fewshot** configurations performed effectively in the binary classification task of screening titles and abstracts, aligning with the goal of maximising sensitivity for inclusion without compromising balanced accuracy in the early phase of SLRs (Figure 5). These results are consistent with previous studies reporting promising performance of LLMs in screening tasks (Ghossein et al., 2025; Syriani et al., 2024; Zuo et al., 2025). Our study extends prior work by combining validated best practices from multiple prior studies and comparing a large ‘over-the-counter’ configuration (GPT-5 Zeroshot) with a smaller, ‘prescribed’ configuration (GPT-4o-mini Fewshot). Findings of the study have several methodological implications.

For researchers less experienced with LLM configuration and prompt engineering, **GPT-5 Zeroshot** offers a convenient, ready-to-use solution. Its larger reasoning capacity reduces the burden of identifying relevant borderline examples for prompts and customised parameters. However, this benefit comes at the cost of higher computational expense (approximately 16 times) and longer task completion time (approximately 4 times) than its counterpart. By contrast, **GPT-4o-mini Fewshot** allows researchers to achieve the same level of accuracy with greater efficiency, offering lower costs and faster throughput. However, this configuration requires fine-grained Fewshot prompt engineering.

The error analysis (Table 1) highlights areas in which refined model instructions and review protocols can improve screening performance. First, ambiguities in the inclusion and exclusion criteria (ET1) can be mitigated by revising eligibility rules to account for borderline cases, noted as a challenge in prior LLM-based screening studies (Matsui et al., 2024; Nykvist et al., 2025). Second, misclassification of articles using documentary or secondary sources (e.g., syllabi, policies, curricula) as relevant can be addressed by clearly instructing the models to differentiate such materials from empirical data generated through student or faculty participation (ET2), addressing prior concerns about criteria clarity (Syriani et al., 2024). Third, smaller models optimised for

consistent responses (**GPT-4o-mini Fewshot** with temperature = 0 and top_p = 0.4) tended to apply rules too rigidly, resulting in exclusion of relevant studies (ET3). Confidence scores and threshold adjustments may provide a partial remedy: for instance, prompts can be designed to impose a higher bar for exclusion when using smaller models to reduce false negatives (Chibwe et al., 2025; Guo et al., 2024; Issaiy et al., 2024).

From the perspective of human-AI collaboration, several benefits emerged. As reported in prior studies, the integration of models can substantially reduce human workload by automating large portions of the initial screening process (Guo et al., 2024; Zuo et al., 2025). In our study, the high sensitivity to eligible articles achieved by both **GPT-5 Zeroshot** and **GPT-4o-mini Fewshot** (Figure 5) suggests that LLMs can reliably assist researchers by filtering out clearly ineligible records, allowing human reviewers to concentrate on evaluating borderline cases. Notably, we also identified cases in which the gold standard needed to be revised in light of the model's decisions. This highlights the potential of LLMs to serve not only as screening accelerators but also quality assurance tools. These efficiencies, combined with the capacity to uncover human oversights, suggest that LLM-assisted screening can meaningfully enhance the speed and reliability of SLRs.

Conclusion

Our study demonstrates that LLM-assisted screening can reduce researchers' manual screening efforts while preserving high screening quality for SLRs. The error analysis identified methodological guidance for future LLM-assisted screening in SLRs. Lastly, LLMs helped correct human miscoding, highlighting their potential role as quality assurance tools rather than benchmarks against humans.

Thus, the study offers methodological contributions. Overall, the study provided a reusable template for recall-oriented, governed screening of a large, high-noise candidate pool, which is often observed in SLRs (Matsui et al., 2024; Syriani et al., 2024). Specifically, it implemented an auditable screening workflow with calibrated prompts, simple ensembling, and majority voting. It also compared a large zeroshot model with a smaller fewshot model on accuracy, recall, time, and cost and reported bootstrap confidence intervals for key metrics. Finally, reframing LLMs as collaborative agents in evidence synthesis rather than post hoc baselines, the study treated human-LLM disagreements as signals to refine eligibility rules and correct gold standards. However, since the reported performance is based on a relatively small, domain-specific corpus, future work should evaluate these guidelines on larger datasets and in additional domains. This will help assess generalisability and refine best practices.

About the authors

Wonchan Choi is an Associate Professor in the School of Information Studies, University of Wisconsin-Milwaukee, USA. He received his PhD from Florida State University. His research interests include GenAI literacy and credibility assessment of GenAI-generated information. He can be contacted at wchoi@uwm.edu

Joyce Lee is a graduate student at the University of Wisconsin-Milwaukee's School of Information Studies. They can be contacted at jlee52@uwm.edu

Besiki Stvilia is a Professor in the School of Information at Florida State University, USA. He received his PhD from the University of Illinois at Urbana-Champaign. His research interests include data quality assurance, digital data curation, online communities, and social informatics. He can be contacted at bstvilia@fsu.edu

Yan Zhang is a Professor in the School of Information at the University of Texas at Austin, USA. She received her PhD from the University of North Carolina at Chapel Hill. Her research interests

include consumer health informatics, information behaviour, information architecture, and recently GenAI literacy. She can be contacted at yanz@utexas.edu

Hyerin Bak is a Teaching Faculty in the School of Information Studies at University of Wisconsin-Milwaukee. She received her PhD from Florida State University. Her research interests include GenAI literacy and media literacy. She can be contacted at hbak@uwm.edu

References

- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, 46(5), 423–429. [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
- Chibwe, K., Mantilla-Calderon, D., & Ling, F. (2025). Evaluating GPT models for automated literature screening in wastewater-based epidemiology. *ACS Environmental Au*, 5(1), 61–68. <https://doi.org/10.1021/acsenvironau.4c00042>
- Ghossein, J., Hryciw, B. N., Ramsay, T., & Kyeremanteng, K. (2025). The AI reviewer: Evaluating AI's role in citation screening for streamlined systematic reviews. *JMIR Formative Research*, 9, e58366–e58366. <https://doi.org/10.2196/58366>
- Guimarães, N. S., Ferreira, A. J. F., Ribeiro Silva, R. D. C., De Paula, A. A., Lisboa, C. S., Magno, L., Ichiara, M. Y., & Barreto, M. L. (2022). Deduplicating records in systematic reviews: There are free, accurate automated ways to do so. *Journal of Clinical Epidemiology*, 152, 110–115. <https://doi.org/10.1016/j.jclinepi.2022.10.009>
- Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, 26(1), e48996. <https://doi.org/10.2196/48996>
- Issaiy, M., Ghanaati, H., Kolahi, S., Shakiba, M., Jalali, A. H., Zarei, D., Kazemian, S., Avanaki, M. A., & Firouznia, K. (2024). Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Medical Research Methodology*, 24, 78. <https://doi.org/10.1186/s12874-024-02203-8>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Matsui, K., Utsumi, T., Aoki, Y., Maruki, T., Takeshima, M., & Takaesu, Y. (2024). Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. *Journal of Medical Internet Research*, 26, e52758. <https://doi.org/10.2196/52758>
- Nykqvist, B., Macura, B., Xylia, M., & Olsson, E. (2025). Testing the utility of GPT for title and abstract screening in environmental systematic evidence synthesis. *Environmental Evidence*, 14(1), 7. <https://doi.org/10.1186/s13750-025-00360-x>
- Oami, T., Okada, Y., & Nakada, T. (2024). Performance of a large language model in screening citations. *JAMA Network Open*, 7(7), e2420496. <https://doi.org/10.1001/jamanetworkopen.2024.20496>
- Syriani, E., David, I., & Kumar, G. (2024). Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages*, 80, 101287. <https://doi.org/10.1016/j.cola.2024.101287>
- Van Dijk, S. H. B., Brusse-Keizer, M. G. J., Bucsán, C. C., Van Der Palen, J., Doggen, C. J. M., & Lenferink, A. (2023). Artificial intelligence in systematic reviews: Promising when

appropriately used. BMJ Open, 13(7), e072254. <https://doi.org/10.1136/bmjopen-2023-072254>

Zuo, C., Yang, X., Errickson, J., Li, J., Hong, Y., & Wang, R. (2025). AI-assisted evidence screening method for systematic reviews in environmental research: Integrating ChatGPT with domain knowledge. Environmental Evidence, 14, 5. <https://doi.org/10.1186/s13750-025-00358-5>

© CC-BY-NC 4.0 (<https://creativecommons.org/licenses/by-nc/4.0/deed.en>) The Author(s). For more information, see our Open Access Policy (<https://publicera.kb.se/ir/openaccess>).