Automating Literature Screening with Multi-Agent Consensus Seeking: A Preliminary Study

Deliang Wang
Faculty of Education
The University of Hong Kong
Hong Kong SAR
wdeliang@connect.hku.hk

Xian Chen
School of Environment, Education and Development
The University of Manchester
UK
metaxianchen@gmail.com

Jiaying (Joy) Yu
Ontario Institute for Studies in Education
The University of Toronto
Canada
joyy.yu@mail.utoronto.ca

Gaowei Chen
Faculty of Education
The University of Hong Kong
Hong Kong SAR
gwchen@hku.hk

ABSTRACT

The automation of systematic reviews (SRs) with large language models (LLMs) offers substantial potential to reduce the labor-intensive demands of manual screening. Yet, single-agent LLMs are prone to unilateral bias and critical omissions (false negatives), particularly when dealing with the contextualized and often subjective constructs characteristic of educational research. This study proposes a heterogeneous dual-agent, consensus-seeking workflow that emulates human double-blind screening followed by inter-coder reconciliation. Using DeepSeek-V3.2 and Qwen3.5-35B as independent screening agents, the system first conducts separate title–abstract evaluations and then applies an automated negotiation mechanism to resolve disagreements. Tested on a dataset of 2,107 articles extracted from the Web of Science across five SRs, the multi-agent framework successfully outperformed both single-agent baselines in accuracy, recall, and F2 scores. Concurrently, it maintained strong workload reduction, accurately excluding 81% of irrelevant records from human review. Variance analyses further demonstrated that the debate mechanism improved the stability of screening performance across heterogeneous data subsets. Although the absolute performance metrics remain moderate—highlighting the inherent difficulty of fully automating literature screening for complex constructs—these preliminary findings provide promising empirical support. They establish multi-agent collaboration as a more rigorous and reliable paradigm for advancing LLM-assisted evidence synthesis in education.

Keywords

Large Language Models, Multi-Agent Systems, Systematic Review, Literature Screening, Consensus Mechanism

Deliang Wang, Xian Chen, Joy Yu, and Gaowei Chen. Automating Literature Screening with Multi-Agent Consensus Seeking: A Preliminary Study. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 594–599. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040068

1. INTRODUCTION

A literature review, as a scholarly paper, provides an overview of current knowledge on a given topic by summarizing and synthesizing the substantive findings of existing research, as well as highlighting theoretical and methodological contributions [13]. Accordingly, literature reviews serve to establish what is known and unknown about a particular phenomenon and to identify areas where further research is required. Based on differences in their search and analysis characteristics, literature reviews can be classified into several types, including rapid reviews, scoping reviews, systematic reviews, meta-analyses, and umbrella reviews, among others [8].

Among these, systematic reviews are among the most frequently employed approaches. They aim to collate all empirical evidence that meets pre-specified eligibility criteria in order to answer specific, clearly formulated research questions [2]. By employing explicit and reproducible systematic methods—such as the PRISMA framework [15])—systematic reviews seek to minimize bias and provide reliable findings from which conclusions can be drawn and informed decisions made. Due to their transparency, accountability, and replicability, systematic reviews have gained increasing popularity within the field of education.

Typically, systematic reviews involve a standardized set of procedures. Retrieved studies are screened for eligibility using predetermined inclusion and exclusion criteria, a process strictly requiring at least two human reviewers working independently to ensure reliability and mitigate individual bias [16]. However, it has been found that conducting a systematic review takes approximately 67 weeks from registration to publication [3]. In research areas characterized by rapid technological change, such a prolonged timeframe may render an unpublished systematic review outdated. Consequently, researchers have begun exploring automated methods to replace human effort in the labor-intensive screening phases [12].

The emergence of large language models (LLMs) presents promising opportunities to expedite this process. Recently, [6] systematically evaluated 18 different LLMs for title-and-abstract screening, demonstrating their substantial potential to reduce workload. However, their findings also highlighted significant variability among models; relying on a single LLM (a single-agent approach) often yields inconsistent accuracy, risking the exclusion of relevant studies or the over-inclusion of irrelevant ones. Furthermore, assigning a single LLM to conduct the screening fundamentally violates the methodological requirement of having multiple independent reviewers, raising concerns about the dependability and credibility of the automated screening process.

To address the limitations of single-agent approaches, recent advancements in qualitative data analysis have introduced multi-agent systems (MAS) (e.g., [17, 18]. In a multi-agent framework, multiple LLMs are instantiated with distinct roles, allowing them to collaborate, cross-verify, and engage in social behaviors such as debating and consensus-seeking [1, 26]. By simulating human collaborative workflows, multi-agent systems can effectively reduce the inherent biases and hallucinations of individual models, yielding far more robust and rigorous analytical outcomes.

Despite these advancements, there is a lack of empirical research examining the application of multi-agent LLM systems to systematic reviews, particularly within the education domain. It remains unclear whether simulating the human dual-reviewer workflow using heterogeneous LLMs (i.e., different underlying models acting as independent coders) can resolve the inconsistency issues identified in prior studies. Therefore, in this preliminary study, we aim to investigate the effectiveness of leveraging a multi-agent framework to facilitate the title-abstract screening process for educational systematic reviews. Specifically, we employ distinct LLMs to act as independent coders, evaluating their individual screening performance and their ability to reach a consensus through simulated inter-coder discussion, thereby paving the way for a more reliable, automated approach to evidence synthesis.

2. RELATED WORK

2.1 LLMs and Multi-Agent Systems in Qualitative Research

LLMs are increasingly utilized in qualitative data analysis (QDA). Early research explored single LLMs as research assistants, demonstrating their ability to simulate thematic analysis [5] and develop codebooks in educational contexts [25, 11]. While single-agent approaches improve efficiency and clarify coding definitions [25], their performance heavily depends on prompting strategies and construct complexity [11, 22, 21, 23]. Furthermore, they remain susceptible to inherent biases and hallucinations when processing complex logic.

To overcome these limitations, research is shifting toward LLM-based multi-agent systems (MAS). By assigning specific roles to different LLMs, MAS can automate complex workflows. For example, [18] utilized task-specific agents for text summarization and theme extraction, showing that this "division of labor" accelerates QDA and ensures analytical validity by approximating the rigor of human research teams.

Crucially, MAS can simulate human "triangulation" and peer-review processes via built-in collaboration. Introducing "debate" and "reflection" workflows enables agents to exhibit consensus-seeking behaviors and reach agreements through negotiation [26, 4]. In thematic analysis, employing multi-agent collaboration (e.g., independent coder roles) significantly enhances the credibility and dependability of findings compared to single agents [17]. Collectively, these mechanisms are critical for improving the rigor of automated QDA.

2.2 Automating Systematic Reviews via LLM Agents

LLMs also present opportunities to accelerate the highly time-consuming systematic review (SR) process, alleviating the burden of manual screening and data extraction. Despite methodological challenges like model bias and hallucinations, researchers are exploring end-to-end automated systems that integrate screening and bias assessment functionalities.

For title and abstract screening, single LLMs show promise but yield mixed results. While GPT-4 achieved a 100% recall rate in environmental science SRs [14], studies evaluating multiple cutting-edge models in medical SRs noted significant discrepancies in false negative and positive rates, alongside substantial instability in identifying relevant literature [19, 6]. Furthermore, extraction accuracy for complex variables remains highly domain-dependent [9].

Despite this potential, two salient research gaps remain. First, the efficacy of LLMs for SRs in the education domain is underexplored. Unlike objective clinical metrics, educational literature involves highly contextualized and subjective constructs, leaving it unclear whether LLMs can reliably apply complex inclusion/exclusion criteria here. Second, most automated SR research relies on single-agent frameworks. This unilateral decision-making is susceptible to inherent biases and fundamentally violates the SR methodological requirement of having at least two independent reviewers. It remains unknown whether a multi-agent system—simulating human dispute discussion and consensus-seeking through independent LLM coders—can surpass single agents. Consequently, empirical research is needed to validate the feasibility of multi-agent collaborative architectures for title-abstract screening in educational SRs.

3. METHOD

3.1 Data Source

In this preliminary study, we searched Computers & Education—a leading journal in the field of education—for systematic reviews with titles containing the phrase "systematic review". Our initial goal was to replicate the literature search process by adhering to the search strategies described in these systematic reviews. However, we found that many reviews did not specify the exact search date, making replication challenging. Additionally, for some systematic reviews, even when we strictly followed the reported search strategies, our search results did not correspond with those presented

Table 1: Five systematic reviews whose literature search was replicated in our study.

Reference	Topic	# Total studies identified in WoS	# Included WoS studies
[7]	Immersive virtual reality serious games for evacuation training and research	249	9
[10]	Speech-recognition chatbots for language learning	235	49
[20]	Computational thinking through programming in K-12 education	643	59
[24]	Teaching and learning robotics content knowledge in K-12 education	373	6
[27]	Impact of esports participation on the development of 21st century skills in youth	607	49

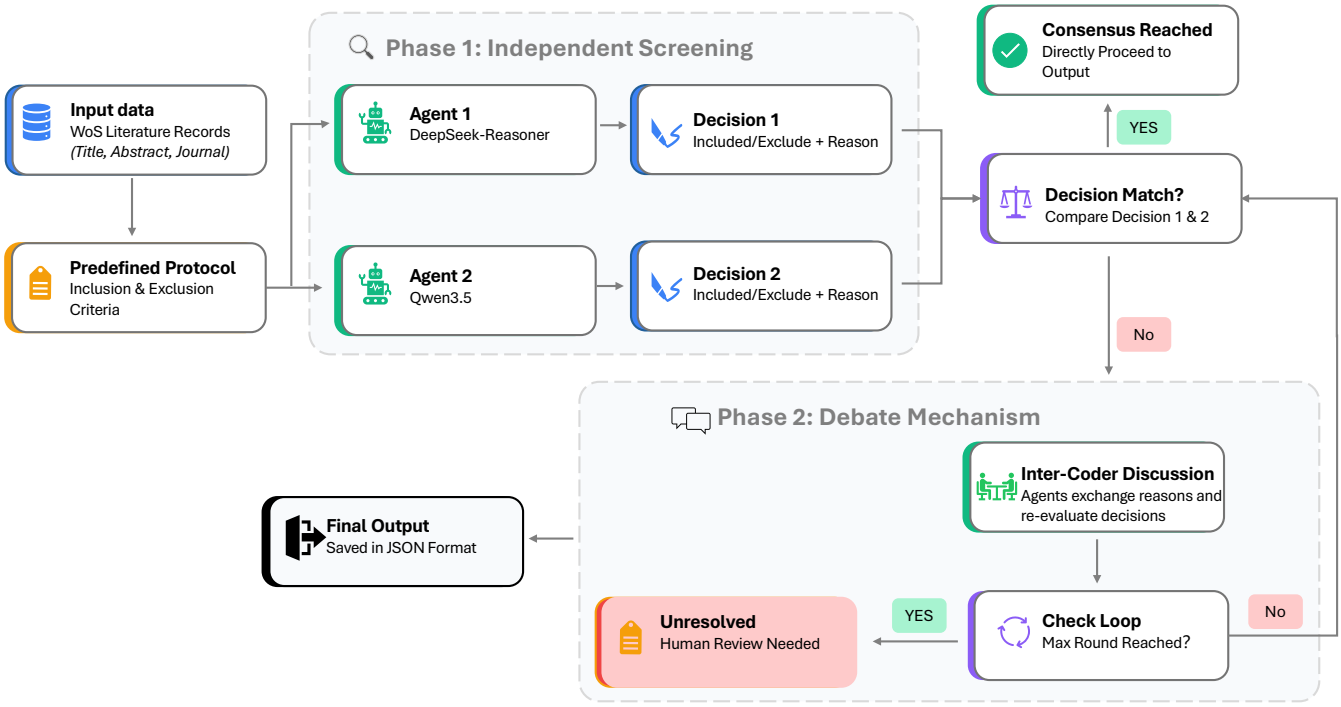

Figure 1: A dual-agent consensus-seeking workflow for systematic literature screening.

in the articles. Consequently, we revised our approach: we limited our analysis to systematic reviews that clearly reported both their search strategies and search dates. We also acknowledged that some variation between our search results and those reported in the articles was likely.

To avoid duplicate records from multiple databases, we restricted our search to the Web of Science (WoS) database. Ultimately, we selected five systematic reviews for this exploratory and preliminary study and replicated their literature search results using WoS. Table 1 summarizes the basic information for these reviews, including their topics, the total number of studies identified in WoS, and the number of WoS studies ultimately included in each review.

3.2 A dual-agent consensus-seeking workflow for systematic literature screening

In this study, we designed a dual-agent consensus-seeking workflow to simulate the rigorous human double-blind screening process for systematic literature reviews, as illustrated in Figure 1. To perform the title-abstract screening, we instantiated two distinct agents powered by heterogeneous large language models (i.e., DeepSeek-V3.2 and Qwen3.5-35b).

In Phase 1, the two agents independently coded each article and generated unilateral decisions regarding its inclusion or exclusion for subsequent full-text review. Both agents were governed by a unified system prompt that assigned them the persona of an expert academic researcher conducting a preliminary systematic review screening. Through zero-shot prompting, the models were provided with strict inclusion and exclusion criteria alongside a standardized output format (e.g., JSON). Importantly, to facilitate an objective evaluation of our workflow's efficacy, these screening criteria were directly extracted from the previously published systematic review. This design choice ensured an unbiased benchmark, allowing us to evaluate the agents' performance

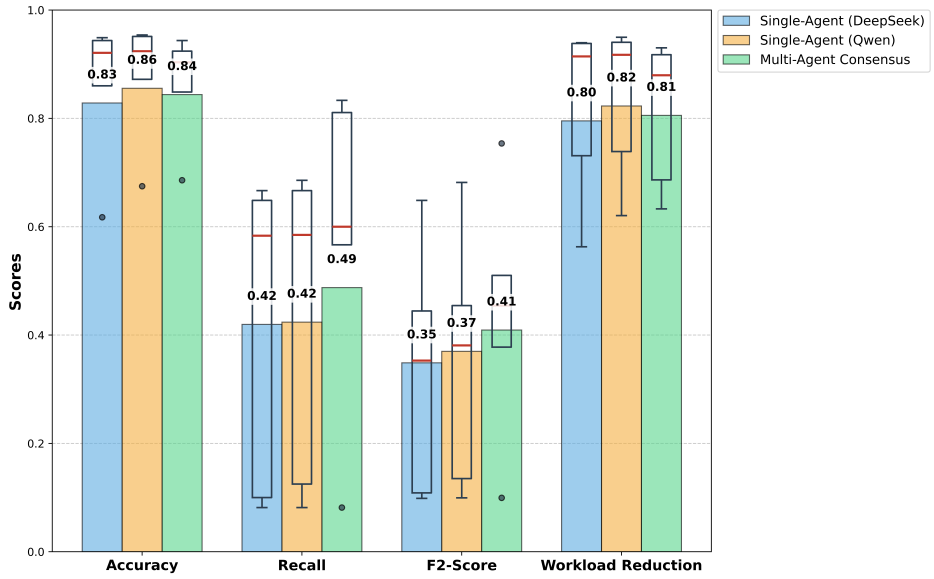

Figure 2: The aggregated global performance (represented by the bar heights) alongside the cross-sectional variance across the five systematic review (illustrated by the overlaid boxplots).

against an established human-coded ground truth.

In Phase 2, the workflow automatically cross-verified the decisions generated by the two agents. If a consensus was reached (i.e., both agents agreed to include or exclude), the final decision was recorded, and the system proceeded to the next article. However, in the event of a disagreement, the debate mechanism was triggered. During a debate round, each agent was provided with the counterpart's previous decision and rationale, prompting them to reflect, cross-examine the discrepancies, and re-evaluate their own stance. If this inter-coder discussion led to a consensus, the loop terminated successfully. If the conflict persisted, the agents engaged in further debate iterations until a predefined maximum number of rounds (i.e., $N = 3$ in our study) was reached. Articles that remained unresolved after the maximum debate rounds were flagged for human intervention by a senior researcher, perfectly mirroring the standard dispute-resolution protocol of traditional systematic reviews.

3.3 Evaluation

To rigorously evaluate the performance of the single-agent baselines (Phase 1) against the multi-agent consensus framework (Phase 2), we compared the classification decisions generated by DeepSeek and Qwen against the human-annotated ground truth (i.e., the ultimate inclusion or exclusion decisions established by the original systematic review).

Given the highly imbalanced nature of title-abstract screening in systematic reviews—where negative samples (excluded articles) vastly outnumber positive ones—we employed four specific evaluation metrics: Recall, F2-Score, Accuracy, and Workload Reduction. Recall calculates the proportion of genuinely relevant studies successfully identified for inclusion. This metric is paramount in our context, as minimizing the omission of relevant literature (i.e., false negatives) is the primary objective of automated screening. To comprehensively evaluate the model's discriminative power while heavily penalizing missed articles, we utilized the F2-Score, which assigns twice the weight to recall compared to precision. Accuracy measures the overall proportion of studies correctly classified by the agents. Finally, Workload Reduction quantifies the practical utility of the automated system by calculating the proportion of total retrieved articles that were correctly excluded (true negatives), reflecting the actual human screening effort saved.

4. PRELIMINARY RESULTS

Due to page limits, Figure 2 reports the aggregated global performance (bar heights) together with the cross-sectional variance across the five independent data chunks (overlaid boxplots).

As shown in Figure 2, the Multi-Agent Consensus system clearly outperformed the single-agent baselines in identifying positive samples. The dual-agent debate mechanism achieved the highest global Recall of 0.49, outperforming both DeepSeek (0.42) and Qwen (0.42). This 16.6% relative improvement suggests that the inter-coder discussion and re-evaluation phase successfully recovered false negatives that were initially missed under unilateral decision-making. Consistent with this pattern, the multi-agent system also obtained the highest F2-score (0.41), compared with DeepSeek (0.35) and Qwen (0.37).

Although Recall was the primary objective, the automated system must also substantially reduce human screening effort. For Workload Reduction, all three configurations performed strongly, effectively filtering out irrelevant records. The Multi-Agent system achieved a Workload Reduction of 0.81 (i.e., safely excluding 81% of all articles from human review), closely comparable to Qwen (0.82) and DeepSeek (0.80). Overall Accuracy ranged from 0.83 to 0.86 across models, indicating a consistently strong ability to identify

true negatives in the highly imbalanced dataset.

The overlaid boxplots in Figure 2 also highlight the intrinsic challenges of automated screening in subjective, imbalanced educational datasets. Accuracy and Workload Reduction exhibited compact interquartile ranges, indicating stable performance across the five data chunks. In contrast, Recall and F2-score displayed larger variability, suggesting that model sensitivity is context-dependent and fluctuates with the specific subset of literature. Notably, however, the Multi-Agent Consensus system raised the median Recall (red line within the box) to approximately 0.60, higher than both single-agent baselines. This upward shift in the median distribution provides additional evidence that the debate mechanism not only improves global performance but also enhances the robustness of screening across heterogeneous samples.

Despite these promising results, 92 studies in the corpus still failed to reach a consensus after three discussion rounds.

5. DISCUSSION

This study proposed a dual-agent consensus-seeking workflow to automate title-abstract screening in educational systematic reviews. Our preliminary findings demonstrate that simulating a human-like inter-coder debate significantly improves global Recall (0.49) and F2-Score compared to single-agent baselines (0.42), while successfully reducing the human screening workload by over 80%. This performance gain highlights the effectiveness of the multi-agent debate mechanism in rectifying unilateral LLM biases and rescuing critical false negatives.

Despite these improvements, the absolute Recall of approximately 50% indicates that fully autonomous LLM screening remains challenging in the education domain. We attribute this to the highly contextualized and subjective nature of educational constructs. Unlike explicit clinical metrics in medical reviews, educational inclusion criteria (e.g., "21st-century skills" or "pedagogical integration") often require nuanced, human-level interpretative reasoning that zero-shot prompting struggles to fully capture. Thus, future iterations should incorporate few-shot prompting with human-annotated examples to calibrate the agents' sensitivity.

6. REFERENCES

[1] A. Barany, N. Nasiar, C. Porter, A. F. Zambrano, A. L. Andres, D. Bright, M. Shah, X. Liu, S. Gao, J. Zhang, et al. Chatgpt for education research: Exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*, pages 134–149. Springer, 2024.

[2] J. Bettany-Saltikov and R. McSherry. How to do a systematic literature review in nursing: A step-by-step guide, 3/e. 2024.

[3] R. Borah, A. W. Brown, P. L. Capers, and K. 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.

[4] H. Chen, W. Ji, L. Xu, and S. Zhao. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*, 2023.

[5] S. De Paoli. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4):997–1019, 2024.

[6] F. M. Delgado-Chaves, M. J. Jennings, A. Atalaia, J. Wolff, R. Horvath, Z. M. Mamdouh, J. Baumbach, and L. Baumbach. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2):e2411962122, 2025.

[7] Z. Feng, V. A. González, R. Amor, R. Lovreglio, and G. Cabrera-Guerrero. Immersive virtual reality serious games for evacuation training and research: A systematic literature review. *Computers & Education*, 127:252–266, 2018.

[8] M. J. Grant and A. Booth. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108, 2009.

[9] T. Jansen, L. W. Liebenow, U. Mertens, F. T. Schmidt, J. F. Lohmann, J. Fleckenstein, and J. Meyer. Data extraction by generative artificial intelligence: Assessing determinants of accuracy using human-extracted data from systematic review databases. *Psychological Bulletin*, 151(10):1280, 2025.

[10] J. Jeon, S. Lee, and H. Choe. Beyond chatgpt: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 206:104898, 2023.

[11] X. Liu, A. F. Zambrano, R. S. Baker, A. Barany, J. Ocumpaugh, J. Zhang, M. Pankiewicz, N. Nasiar, and Z. Wei. Qualitative coding with gpt-4: Where it works better. *Journal of Learning Analytics*, 12(1):169–185, 2025.

[12] I. J. Marshall and B. C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):163, 2019.

[13] M. Newman and D. Gough. Systematic reviews in educational research: Methodology, perspectives and application. *Systematic reviews in educational research: Methodology, perspectives and application*, pages 3–22, 2019.

[14] B. Nykvist, B. Macura, M. Xylia, and E. Olsson. Testing the utility of gpt for title and abstract screening in environmental systematic evidence synthesis. *Environmental Evidence*, 14(1):7, 2025.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

[16] A. Pollock and E. Berge. How to do a systematic review. *International Journal of Stroke*, 13(2):138–156, 2018.

[17] T. Qiao, C. Walker, C. Cunningham, and Y. S. Koh. Thematic-lm: a llm-based multi-agent system for large-scale thematic analysis. In *Proceedings of the*

ACM on Web Conference 2025, pages 649–658, 2025.

[18] Z. Rasheed, M. Waseem, A. Ahmad, K.-K. Kemell, W. Xiaofeng, A. N. Duc, and P. Abrahamsson. Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis. *arXiv preprint arXiv:2402.01386*, 2024.

[19] M. Ruan, J. Fan, M. Liu, Z. Meng, X. Zhang, and C. Zhang. Artificial intelligence for the science of evidence synthesis: how good are ai-powered tools for automatic literature screening? *BMC Medical Research Methodology*, 25(1):199, 2025.

[20] C. Tikva and E. Tambouris. Mapping computational thinking through programming in k-12 education: A conceptual model based on a systematic literature review. *Computers & Education*, 162:104083, 2021.

[21] D. Wang and G. Chen. Evaluating the use of bert and llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy. *British Journal of Educational Technology*, 56(6):2671–2704, 2025.

[22] D. Wang, D. Shan, Y. Zheng, K. Guo, G. Chen, and Y. Lu. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th international conference on educational data mining*, pages 515–519, 2023.

[23] D. Wang, Y. Zheng, J. Li, and G. Chen. Parameter-efficiently fine-tuning large language models for classroom dialogue analysis. *IEEE Transactions on Learning Technologies*, 18:542–555, 2025.

[24] L. Xia and B. Zhong. A systematic review on teaching and learning robotics content knowledge in k-12. *Computers & Education*, 127:267–282, 2018.

[25] A. F. Zambrano, X. Liu, A. Barany, R. S. Baker, J. Kim, and N. Nasiar. From ncoder to chatgpt: From automated coding to refining human coding. In *International conference on quantitative ethnography*, pages 470–485. Springer, 2023.

[26] J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, and S. Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

[27] Y. Zhong, K. Guo, J. Su, and S. K. W. Chu. The impact of esports participation on the development of 21st century skills in youth: A systematic review. *Computers & Education*, 191:104640, 2022.