AI- and Computer-Assisted Systematic Literature Reviews WIP

Jorge Andrés Cristancho Rodríguez, Purdue Engineering Education

Hola, my name is Jorge Cristancho Rodríguez. I am an Engineering Education PhD candidate at Purdue University. I earned my BS and master's in electrical and computer engineering from Los Andes University. My research interests include cultivating eco-social values in engineering education, teamwork, ethics, computer- AI-assisted learning, and student-centered, personalized, and liberatory pedagogies.

Prof. Alejandra J. Magana, Purdue University

Alejandra J. Magana, Ph.D., is the W.C. Furnas Professor in Enterprise Excellence of Applied and Creative Computing and Professor of Engineering Education at Purdue University

Miguel Alfonso Feijoo-Garcia, Purdue Polytechnic Institute, Purdue University – West Lafayette

I am a Ph.D. Candidate in Computer and Information Technology at the School of Applied and Creative Computing at the Polytechnic Institute of Purdue University, West Lafayette. My research interests focus on applied analytics to support computing and engineering education. I am currently working in the Research of Computing in Engineering and Technology Education Lab (RocketEd) under Dr. Alejandra J. Magana. I am from Colombia, where I earned a B.Sc. in Systems and Computing Engineering, a B.Sc. in Environmental Engineering, an M.Sc. in Information Engineering, and an M.Ed. in Higher Teaching Education.

WIP Methods for Computer-Assisted Systematic Literature Reviews in STEM

Jorge Cristancho Rodríguez, Miguel Feijoo, Alejandra Magana. Purdue University.

Abstract
Systematic Literature Reviews (SLRs) are essential for synthesizing scientific knowledge. One of the main challenges of SLRs is the substantial time required for repetitive tasks, such as data collection and screening based on inclusion/exclusion criteria aligned with the research question(s). Recent advances in Machine Learning (ML) and Large Language Models (LLMs) provide new opportunities to support researchers throughout the SLR process. In this study, ML and LLMs were integrated across the main stages of the review to refine searches, assist screening, facilitate data collection, and support synthesis. This Work in Progress specifically explores the integration of computer-assisted methods into the screening phase of SLRs, following the PRISMA framework. The study highlights emerging practices and addresses eco-social ethical considerations at each stage, including transparency, energy consumption, potential declines in researchers' skills, and data privacy. We argue that computational tools do not replace human expertise, but rather enhance the scalability of analysis when guided by methodological structure, human verification, and iterative processes. Ultimately, the findings aim to offer a set of good practices for the responsible, ethical, and effective use of computer- and AI-assisted methods in SLRs.

Keywords: SLR, PRISMA, computer-assisted, AI-assisted, machine learning, large language models

1. Introduction

In this work in progress, we present how machine learning (ML) and large language models (LLMs) can assist researchers in conducting the screening phase of systematic literature reviews (SLRs) in accordance with the PRISMA framework. We also address ethical considerations that ought to be taken into account during this process. This work is significant for two main reasons: pragmatism and ethics. Computational tools have transformed how we conduct research, offering capabilities that extend far beyond what human researchers can accomplish alone in terms of speed and scale of information processing. When implemented thoughtfully, these tools can reduce the mechanical labor inherent in systematic reviews, such as the initial screening of thousands of abstracts or the syntactic verification of inclusion/exclusion criteria, allowing researchers to address larger problems and focus on higher-order tasks, including ideation, problem-solving, critical analysis, and decision-making.

These considerations should include questions of power, learning, and impact, such as authorship and intellectual credit, environmental impact (e.g., energy costs), training of essential research skills (e.g., writing and discerning), and global disparities in countries' economic development (e.g., global south creators with no authorship protection). We use an SLR to show the pros and cons of using computer-assisted tools (i.e., ML and LLMs), and for this WIP, we focus on the screening phase of the PRISMA process. By examining pragmatic and ethical dimensions of computer-assisted SLRs, we aim to contribute to a more responsible integration of powerful tools into research practice, such as SLRs.

2. Background

SLRs are essential for synthesizing scientific knowledge across disciplines, as they employ explicit, reproducible methods to identify, appraise, and synthesize the most relevant studies on a topic, thus clarifying evidence (Martínez et al., 2025). SLRs also serve as the foundation of evidence-based decision-making, underpinning meta-analysis, theory-building, guideline development, and policy decisions (De Cassai et al., 2025; Whittemore et al., 2014). This methodology distinguishes SLRs from narrative reviews by producing reproducible summaries that inform research, practice, and policy. SLRs further conglomerate and interpret bodies of primary studies to resolve inconsistencies, estimate effect sizes, and generate higher-level conclusions that single studies cannot provide (De Cassai et al., 2025; Whittemore et al., 2014). Quantitative synthesis combines effect estimates using meta-analysis, while qualitative synthesis integrates textual data to produce context-sensitive conceptual models. They also identify evidence gaps, guiding future research priorities and funding allocations.

Nevertheless, SLRs face substantial practical and methodological challenges. For instance, the process is resource-intensive, often requiring months or years to complete (Okoli, 2015). In particular, the screening phase, in which researchers manually examine hundreds to thousands of titles and abstracts, is time-consuming. Rathbone et al. (2017) found that citation screening alone can take days or months. It is also error-prone, with documented human error rates between 6% and 21%, meaning that even with dual independent screening, relevant studies may be inadvertently excluded (Wang et al., 2020; Pang et al., 2025).

To address these challenges, researchers have increasingly explored computational tools, such as ML and LLMs, to automate stages of the systematic review process (Zhang et al., 2022). Machine learning techniques have evolved from simple keyword-based filtering to sophisticated classification systems, including automated classification with human-in-the-loop feedback (Mosqueira-Rey et al., 2024). These methods can reduce workload while maintaining acceptable recall, even though LLMs are non-deterministic algorithms.

The emergence of large language models has further increased interest in automating systematic reviews. Platforms such as SPECTRUM SRA have integrated LLMs into comprehensive workflows, achieving substantial inter-model agreement for screening tasks (Nezhad et al., 2025). A scoping review by Lieberum et al. (2025) examined 37 studies on LLM applications in systematic reviews and found that LLMs covered 10 of 13 defined review steps, most often for literature search (41%), study selection (38%), and data extraction (30%). However, the evaluations were mixed: twenty studies found LLMs promising (54%), nine neutral (24%), and eight non-promising (22%).

At the same time, algorithmic bias in AI-assisted systematic reviews is an emerging concern. Pardal-Refoyo and Pardal-Peláez (2025) found that while frameworks exist to address bias in clinical AI, significant gaps remain in applying these frameworks to AI-assisted systematic reviews. This highlights the need for future research to ensure transparency, fairness, and methodological consistency in evidence synthesis.

3. Methods

This work in progress focuses on the screening phase of the PRISMA framework. **Appendix A**: *PRISMA figure* **Appendix A**: *PRISMA* illustrates all four phases of the PRISMA framework and highlights the checklist items potentially supported by machine assistance: Phase 1: *Identification* (Items 1-2), Phase 2: *Screening* (Items 9, 12, and 15), and Phase 3: *Eligibility Assessment* (Items 17, 19, and 22). Items outside this scope are addressed descriptively by phase. The figure only shows the PRISMA elements that a machine can/should support. The remaining items are addressed descriptively in this work according to their corresponding phase. That is, as this work in progress focuses on Phase 2: *Screening*, it is therefore emphasized in **Appendix A**: *PRISMA figure* and in this section. This work in progress followed the following process: *Inclusion-exclusion criteria:* Two researchers analyzed 50 abstracts together and created an inclusion/exclusion framework (**Appendix B**: *Inclusion exclusion framework*). *Human screening:* using the framework, one researcher analyzed half of the abstracts, and another researcher analyzed the other half. Each researcher classified the abstracts as included, excluded, or unclear. Then, the two researchers jointly resolved all unclear abstracts, producing a human-coded baseline. Afterwards, we used ML, LLMs, and agentic LLMs to compare our baseline with their results.

Machine learning screening: For the machine learning screening process, we used two strategies, syntactic and semantic analysis. First, we conducted a syntactic examination of all abstracts, focusing on the presence and structure of words. Then, we performed a semantic analysis supported by the previous syntactic analysis by comparing the relationships between words in sentences to understand the grammatical context. We used Latent Semantic Analysis (LSA) to better understand the meaning of words through their relationships. LSA is a technique in natural language processing that analyzes relationships between words across multiple documents by producing a set of concepts related to the documents and terms (Deerwester et al., 1990; Landauer et al., 1998).

Large language model screening: We compared the performance of four different LLMs. In all cases, we used the following protocol: **a)** used a single-turn prompt, that is, we did not reuse threads or previous conversations; **b)** used an identical prompt across all models (First part of **Appendix E**: *LLM and Agentic AI prompts*); **c)** fixed sampling parameters (e.g., temperature=0, top_p=1); **d)** used the same model in all the iterations (e.g., ChatGPT 5.3); **e)** Normalized output by requesting a fixed 6 by 662 table in Excel; **f)** we ran 5 trails with each LLM; **g)** we controlled for system prompt by asking the LLM to be a SLR reviewer, to be consistent, to only use information provided in the input documents, to show the code snippets used in the process, and to show all evidence of tentative decisions; **h)** finally, we logged all prompts, response time, outputs, model version, and parameters.

Agentic LLM screening: For the Agentic LLM, we used paid ChatGPT 5.3 and Claude Sonnete 4.6. We used the same conditions as for the LLM screening, plus we added the second part of **Appendix E**: *LLM and Agentic AI prompts*, and the system prompts that it could decide to look compare with other abstracts, look for information anywhere in the input files, including PRISMA references online, improve the framework, and report all decisions that changed the command. The only reported change was the addition of a seventh column providing reasons for exclusion based on the criteria provided in the **Appendix B**: *Inclusion exclusion framework*.

4. Preliminary results

We started with 1,087 articles from four databases (see Appendix F: Database search protocol). We then reduced them to 671 after removing duplicates (413) and retracted articles (3). We ran the syntactic exclusion and found 9 articles that did not include our keywords. After human revision, we confirmed these 9 articles were 100% accurately excluded. This left us with 662 abstracts to screen.

Using the **Appendix B**: *Inclusion exclusion framework* one researcher screened 331 abstracts, and another screened the other 331. We had 18 unclear abstracts, which we resolved after revising the framework and discussing our perspectives. With that, we excluded 589 abstracts and ended up with 187 abstracts to include in the full-text literature retrieval. This 662-list of included or excluded abstracts became the baseline for the comparison.

For the Machine Learning test, we used syntactic and semantic analysis, specifically, we used Latent Semantic Analysis to compare with our inclusion and exclusion decisions. The ML algorithm had an error of 29.15%. That is, human and LSA had an agreement of 70.85% (469/662).

For the next step of this comparative analysis, we replicated the screening process using four LLMs (i.e., Claude, ChatGPT, Gemini, and Grok) following the same inclusion/exclusion framework presented in **Appendix B**: *Inclusion exclusion framework*. Only Claude (Sonnet 4.6) and ChatGPT (GPT 5.3) were able to produce an outcome with the decision of inclusion or exclusion for each of the 662 abstracts. After running the prompt 5 times, the average response time of Claude was 92.9 seconds with an error of 43.05%, and the average response time of ChatGPT was 35.4 seconds with an error of 35.20%.

The fact that LLMs are non-deterministic led us to create agentic prompts that guided step-by-step the process of screening (see **Appendix E**: *LLM and Agentic AI prompts*). Following eight steps, Claude had an average error of 36.01% in around 141.2 seconds using the extended thinking feature. In contrast, ChatGPT improved the error to 28.43% with an average response time of 34.52 seconds. See Table 1. Screening comparison for the full summary.

Table 1. Screening comparison

Type of Machine	Error	Time (seconds)
ML	29.15%	21.36
ChatGPT (5.3)	35.20%	34.4
Claude (Sonnet 4.6)	43.05%	92.9
Agentic ChatGPT	28.43%	34.52
Agentic Claude	36.01%	141.2

5. Ethical considerations: social and environmental

The integration of computer-assisted models into research practices requires careful consideration of their algorithmic implications on society and Nature. These models can perpetuate discrimination through a false appearance of neutrality when algorithms are trained with biased data. Pardal-Refoyo and Pardal-Peláez (2025) argue that algorithmic bias appears across five pipeline stages: data collection, preparation and annotation, model development, evaluation, and implementation, demonstrating the pervasive nature of this challenge.

Similarly, Ruha Benjamin (2019) highlighted this bias problem by introducing the "New Jim Code" to describe how new technologies reproduce existing inequities while being promoted as objective or neutral. She argues that when algorithms are trained on biased historical data, they actively perpetuate, reinforce, and amplify human bias, creating "engineered inequity," a dangerous feedback loop in which past discrimination becomes normalized and automated, hidden behind apparent computational neutrality. Benjamin notes that the main danger is assuming technology is neutral and objective, which leads us to question it far less than human decision-makers.

A second social impact concerns the atrophy of essential academic skills. Writing is a form of thinking and learning, and drafting and revising help develop conceptual clarity, analytical reasoning, and argumentative coherence. When researchers outsource substantial writing to LLMs, particularly in formative stages, they risk losing these capacities. A third impact involves appropriating freely contributed work for commercial gain. Since 2001, millions of volunteers have created Wikipedia content for this nonprofit repository, yet LLM developers have disrupted this social contract by scraping open content into commercial products.

Researchers should also consider the environmental implications of LLMs, which require substantial energy for training and use. Training consumes significant electricity, produces carbon emissions, and requires water and electronic resources (Mu et al., 2025; Elsworth et al., 2025). While individual queries consume little energy, the cumulative impact is substantial given billions of daily queries. Advanced models can consume over 33 Wh per long prompt (Elsworth et al., 2025), and data centers require additional resources (Mu et al., 2025). To reduce these impacts, researchers can use Retrieval-Augmented Generation (RAG), which reduces retraining needs (Zhao et al., 2026; Han et al., 2024), or smaller language models (SLMs) focused on specific tasks with less computing power (Wang et al., 2025). Combining SLMs with RAG can lower energy consumption while maintaining performance (Wu et al., 2022; Verdecchia et al., 2023). Furthermore, researchers must always explicitly report on the use of LLMs in their work, detailing which tools were employed and for what purposes. Ideas should originate from researchers, and research decisions should be guided by humans with a critical perspective on power structures and eco-social implications.

6. Limitations, future work, and conclusions

Computer-assisted Systematic Literature Reviews have different advantages, especially through the use of Machine Learning and AI-driven technologies. It also presents important limitations. Its effectiveness depends on human knowledge, as researchers must understand systematic review methods, how the algorithms operate, and have basic knowledge of the reviewed topic. Consequently, outputs must be interpreted critically, and errors, such as hallucinations, must be corrected, making iterative human-in-the-loop processes essential. Even experienced researchers cannot rely on these tools alone because all decisions require careful verification, and machines cannot replace human judgment or independently generate reviews.

Another challenge is the demand for computational resources, including processing power, storage, and cloud access, which can limit the ability to handle large datasets or run complex algorithms. Access to full-text articles is also often restricted by paywalls, reducing review completeness and creating gaps. In addition, these technologies may introduce algorithmic bias, favoring some studies while overlooking others. Human oversight is therefore always necessary

to detect and correct such biases, ensuring the integrity of the review process. In this way, researchers must verify all AI-generated content, balancing efficiency and accuracy while preserving methodological rigor, ethical integrity, and the quality of human contributions in systematic literature reviews.

Future work should focus on advancing agentic AI processes and evaluating error reduction through fine-tuned LLMs, Small Language Models, or Retrieval-Augmented Generation (RAG) specifically designed for SLRs. Equally important is the development of ethical frameworks to guide researchers in using computer-assisted tools in eco-socially responsible ways. This work demonstrates that SLRs can be supported by machine learning algorithms and LLMs by reducing mechanical labor and increasing replicability through trained verification. At the same time, our findings resonate with Lieberum et al. (2025), who suggest that computational tools are not yet ready to fully implement SLRs. Humans should make the decisions and verify computed contributions with a critical perspective that accounts for quality and eco-social impact.

References

Longworth, J. (2021). Benjamin Ruha (2019) Race After Technology: Abolitionist Tools for the New Jim Code. Medford: Polity Press. 172 pages. eISBN: 9781509526437. *Science & Technology Studies*, *34*(2), 92-94.

De Cassai, A., Dost, B., Tulgar, S., & Boscolo, A. (2025). Methodological standards for conducting high-quality systematic reviews. *Biology*, *14*(8), 973.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

Ding, Z., Wang, J., Song, Y., Zheng, X., He, G., Chen, X., ... & Song, J. (2025). Tracking the carbon footprint of global generative artificial intelligence. *The Innovation*, *6*(5).

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. *arXiv preprint arXiv:2508.15734*.

Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., & Strubell, E. (2025, July). Energy considerations of large language model inference and efficiency optimizations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 32556-32569).

Google DeepMind. (2025, August). Gemini environmental impact report. https://ai.google/responsibility/environmental-impact/

Hacker, P., Mittelstadt, B., Hammer, S., & Engel, A. (2025). Introduction to the foundations and regulation of generative AI.

Han, B., Susnjak, T., & Mathrani, A. (2024). Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview. *Applied Sciences*, *14*(19), 9103.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Lieberum, J. L., Toews, M., Metzendorf, M. I., Heilmeyer, F., Siemens, W., Haverkamp, C., ... & Eisele-Metzger, A. (2025). Large language models for conducting systematic reviews: on the

rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology*, *181*, 111746.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005-3054.

Mu, H., Lu, Y., Lepre, C., Lightfoot, C., Smiley, C., & Crenshaw, R. (2025, May). Artificial intelligence and its carbon footprint. In *2025 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 31-35). IEEE.

Nezhad, H. M., & CheshmehSohrabi, M. (2025). SPECTRUM SRA: An R Shiny Application to Automate Systematic Reviews Using Artificial Intelligence and Large Language Models.

O'Donnell, J., & Crownhart, C. (2025). We did the math on AI's energy footprint. Here's the story you haven't heard. *MIT Technology Review*.

Pang, X., Saif-Ur-Rahman, K. M., Berhane, S., Yao, X., Kothari, K., Taneri, P. E., ... & Devane, D. (2025). Comparing Artificial Intelligence and manual methods in systematic review processes: protocol for a systematic review. *Journal of Clinical Epidemiology*, *181*, 111738.

Pardal-Refoyo, J. L., & Pardal-Pelaéz, B. (2025). Stage-wise algorithmic bias, its reporting, and relation to classical systematic review biases in AI-based automated screening in health sciences: A structured literature review. *medRxiv*, 2025-05.

Rathbone, J., Albarqouni, L., Bakhit, M., Beller, E., Byambasuren, O., Hoffmann, T., ... & Glasziou, P. (2017). Expediting citation screening using PICo-based title-only screening for identifying studies in scoping searches and rapid reviews. *Systematic reviews*, *6*(1), 233.

Ritchie, H. (2025, August 21). What's the carbon footprint of using ChatGPT or Gemini? [August 2025 update]. Hannah Ritchie's Substack. https://hannahritchie.substack.com/p/ai-footprint-august-2025

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, *36*, 8634-8652.

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(4), e1507.

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., ... & Wang, S. (2025). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, *16*(6), 1-87.

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS one*, *15*(1), e0227742.

Wikipedia. (n.d.). About Wikipedia. Retrieved January 20, 2026, from https://en.wikipedia.org/wiki/Wikipedia:About

Whittemore, R., Chao, A., Jang, M., Minges, K. E., & Park, C. (2014). Methods for knowledge synthesis: an overview. *Heart & Lung*, *43*(5), 453-461.

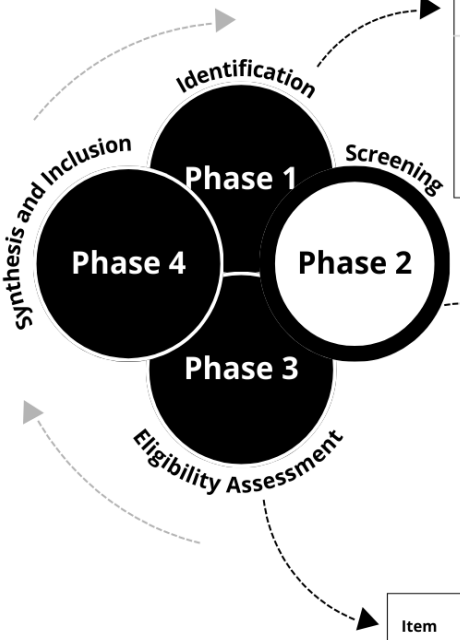
Zhang, Q., Fang, C., Xie, Y., Ma, Y., Sun, W., Yang, Y., & Chen, Z. (2024). A systematic literature review on large language models for automated program repair. *ACM Transactions on Software Engineering and Methodology*.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2026). Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, 1-29.

APPENDICES

Appendix A: *PRISMA figure*

Figure 1 PRISMA phases highlighting the items in the checklist that are potentially supported by machine assistance



Item	PRISMA	Machine support
1	Identify the report as a systematic review	***Name suggestion. Machines can be useful at suggesting alternatives based on topics or the target audience.***
2 (Abstract)	Provide a structured summary that includes background, objectives, data sources, eligibility criteria, participants and interventions, appraisal and synthesis methods, results, limitations, conclusions and implications, and the systematic review registration number.	***Writing assistant: see section below.***

Item	PRISMA	Machine support
9 (Methods)	Study selection: State the process for selecting studies (e.g., Screening)	***Screening: See section below. This is the main focus of this WIP.***
12 (Methods)	Risk of bias in individual studies: Describe methods used for assessing risk of bias of individual studies, and how this information is to be used in any data synthesis.	***See section algorithms are biased.***
15 (Methods)	Risk of bias across studies: Specify any assessment of risk of bias that may affect the cumulative evidence.	***See section algorithms are biased.***

Item	PRISMA	Machine support
17 (Results)	Study selection: Give numbers of studies screened, assessed for eligibility, and included in the review, with the reasons for exclusions at each stage, ideally with a flow diagram.	***See Appendix B***
19 (Results)	Risk of bias within studies: Present data on the risk of bias of each study and, if available, any outcome-level assessment.	***See section algorithms are biased.***
22 (Results)	Risk of bias across studies: Present results of any assessment of risk of bias across studies.	***See section algorithms are biased.***

Appendix B: *Inclusion exclusion framework*

Exclude if:

1. The document studies Students in K-12, graduate school, or graduates, or faculty

2. The document studies faculty, instructors, or professors
3. The study is developed in a laboratory and not a classroom.
4. It is a Review or meta-analysis of teamwork pedagogies and not an application of a teamwork pedagogy
5. Does not focus on engineering
6. Not between 2019 and 2024
7. Check if teams are formed between students and not institutions

Appendix C: *LLM table*

	Claude	**ChatGPT**	**Gemini**	**Grok**
Task completed	Yes	Yes	No	No
Time	92.9	35.4	155.4	39
Version	Sonnet 4.6	GPT 5.3	Gemini 3 Flash	Expert
Included	401	129	23	91
Excluded	261	533	574	530

Table 3. LLM performance

Appendix D: *Computational energy consumption*

Task Type	**Traditional ML Energy consumption**	**LLM Energy Consumption**	**Notes**
Text prediction classification	0.0003-0.03 Wh (1-100J)	0.24-0.43 Wh (short query)	ML models like BERT small: ~114J per query; LLMs 70x higher
Long text generation	Not applicable	1-3Wh	Traditional ML not designed for extended text generation
Image classification	0.0003-0.001 Wh (1-4J)	Not applicable	Traditional CNNs are vastly more efficient for classification
Image generation	Not applicable	~3Wh (standard quality)	Traditional ML cannot generate complex images
Video generation	Not applicable	10+ Wh (can exceed 900Wh)	Traditional ML cannot generate video content
Model training	Varies widely by task	1,000+ MWh (GPT-3: 1,287 MWh)	LLMs require orders of magnitude higher

Appendix E: *LLM and Agentic AI prompts*

Screening protocol:

Please revise this document to provide a table deciding if each of these abstracts should be included or not. The keywords for these documents are: teamwork, “team work”, groupwork, and “group work”; in the context of engineering using the second key words as engineer*, STEM, STM, STEEM; for the space of higher education giving us the third key words “higher ed*”, university, college, and “pedago*” and “instruct*”.

The **inclusion/exclusion framework** is the following:

Exclude if:

1. The document studies Students in K-12, graduate school, or graduates, or faculty
2. The document studies faculty, instructors, or professors
3. performs Laboratory work
4. It is a Review or meta-analysis of teamwork
5. Does not focus on engineering
6. Not between 2019 and 2024
7. Check if teams are formed between students

The output should be a *csv* or *xlx* file with the same input table, with an extra column that indicates *1* if the abstract satisfies these conditions and 0 if it does not.

Agentic only section:

To do this, you are going to do the following steps:

1. Get familiar with the PRISMA process: https://www.prisma-statement.org/prisma-2020-flow-diagram
2. Read structure,
3. Read content
4. Read sample of abstracts
5. Use these examples of exclusion as guidance (Examples of exclusion: 9 excluded because of criteria 7, teams are not formed between students; 97 & 233 excluded because of criteria 2, it focuses on teachers' perspectives; 103 excluded because of criteria 7, the collaboration is between institutions not students. 350 excluded because of criteria 1, focused on 12 graders.)
6. Analyze abstracts with framework
7. Verify inclusion/exclusion with framework and samples
8. Create report .xlsx document with the inclusion column

Appendix F: Database search protocol