Rev. Spirat. 2025;3(SN1): e5639
DOI: 10.20453/spirat.v3iNE1.5639

Can Large Language Models Automate Systematic Studies of Literature? Exploring Automated Screening - A Case Study in the Field of Computer Science Education

Franklin L. Sánchez[1] [0000-0003-3963-2630]
Carlos Alario-Hoyos[2] [0000-0002-3082-0814]
Danny S. Guamán[3] [0000-0003-2794-3079]
Julio C. Caiza[4] [0000-0001-9910-582X]
[1,2] *Universidad Carlos III de Madrid*
[3,4] *Escuela Politécnica Nacional*
franklin.sanchez@alumnos.uc3m.es

Abstract

This study evaluates the efficacy of Large Language Models in the screening process of Systematic Literature Studies in Computer Science Education, a domain with increasing contributions. Using models such as GPT-4o, Claude-3.5-Sonnet, and Llama-3-70B, the automation of the screening process is explored, comparing its results with a manual process carried out by researchers in the area. The data worked with are from July 2024 and the results of the selection process show high sensitivity (≥0.8644) in all models, indicating that at least 86% of the relevant articles are included, and it is highlighted that Claude-3.5Sonnet includes 96.6% of the relevant articles. The F1-Score values for Claude3.5-Sonnet and GPT-4o (≥0.74) show that the models' performance is acceptable for this study's context. Although the low precision (≥0.355) indicates that the models tend to include non-relevant articles, the results obtained suggest that LLMs have significant potential as support tools at the inclusion/exclusion stage, potentially reducing manual review time. However, a hybrid approach combining automation with human judgment in the final tasks of this stage is recommended.

Keywords: Computer Science Education, Large Language Models, Systematic Literature Studies

Introduction

The body of knowledge (BoK) in Computer Science (CS) education has grown exponentially, evidenced by the proliferation of publications in areas such as computational thinking (Muñoz et al., 2020), pedagogical strategies in programming (Medeiros et al., 2019) and the formation of computer-assisted learning groups (Oliveira et al., 2019). The application of artificial intelligence in education has further accelerated this growth, generating a continuous flow of innovations (Liang et al., 2024).

Systematic Literature Studies (SLS) are crucial for organizing this vast BoK, determining the maturity of contributions, and synthesizing emerging trends. However, developing an SLS typically requires 12 to 18 months (Sachs, 2018), delaying the application of new knowledge in real educational scenarios and creating a gap between knowledge generation and its practical implementation.

Large Language Models (LLMs) have emerged as promising tools for automating phases of SLEs. Previous research, such as Castillo et al. in, (Castillo-Segura et al., 2023) has explored their use in Systematic Literature Reviews, with encouraging but improvable results.

Our proposal is distinguished by using state-of-the-art models such as GPT-4o (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024) , and Llama-3-70B (Meta, 2024), known for their superior performance at the time of the elaboration of this study (Hugging Face, n.d.), focusing on automating the screening process of articles in the context of an ESL in the area of CC teaching.

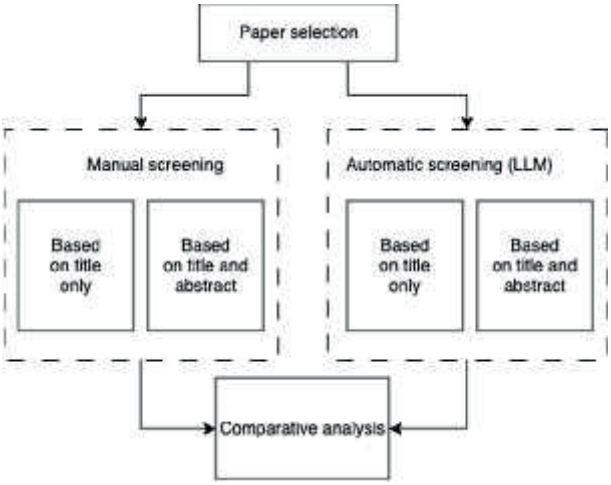
Background

According to Petersen et al. (Petersen et al., 2015), SLEs consist of three main phases: planning, execution, and reporting. The execution phase defines the stages of applying the article selection strategy, applying inclusion/exclusion criteria, classifying articles, analyzing articles, and identifying gaps and trends.

Method

This study implemented the paper selection and screening phases of an ESL, in the field of Computer Science teaching in higher education, following the guidelines proposed in (Petersen et al., 2015), as illustrated in Figure 1.

Figure 1. Method for screening articles



The article selection phase was carried using the Scopus database, and a total of 1055 articles were obtained for subsequent evaluation.

The screening phase was carried out in two stages: evaluation based on titles and analysis of titles and abstracts. This process was carried out in parallel, both manually by three expert researchers and in an automated way using the selected Large Language Models (LLM).

The manual screening involved 3 researchers who applied predefined inclusion/exclusion criteria. One investigator reviewed each article, while the other 2 conducted test pilots to ensure the criteria were rigorously applied.

The three best-performing LLMs were used for the automated screening: GPT-4o, Claude-3.5-Sonnet, and Llama-3-70B. A Python script was developed that interacts with the APIs of these models, and

prompts were built based on a template (Shin et al., 2020) that incorporates some of the common components defined in (Schulhoff et al., 2024), applying the zero-shot prompting technique (Kojima et al., 2022). For screening based on title, the prompt defined in Table 1 was used, while for classification based on title and abstract, a specific prompt was created and applied for each inclusion criterion established in the ESL, maintaining a consistent structure with the initial prompt.

Table 1 Prompt for screening based on title

Table 1 Prompt for screening based on title

Prompt
You are a researcher conducting a systematic literature review. According to the title of this article. Title: {titulo}. Do you consider that the article is related to the use of LLMs in the teaching-learning process of computer science? If you consider yes, answer with 'Included', if you consider no, answer with 'Excluded'.

You are a researcher conducting a systematic literature review. According to the title of this article. Title: {titulo}. Do you consider that the article is related to the use of LLMs in the teaching-learning process of computer science? If you consider yes, answer with 'Included', if you consider no, answer with 'Excluded'.

In order to assess the performance of the models in the screening process, we constructed confusion matrices and calculated accuracy metrics such as recall, precision, and F1-Score. These metrics facilitated a systematic comparison between the effectiveness of LLMs and human investigators, offering a quantitative foundation for evaluating the potential for automation in the screening process.

Results

Table 2 shows the results for the classification based on title, while Table 3 presents the results for classifying articles based on title and abstract.

Tabla 2 Accuracy metrics for title-based screening.

Model	Precision	Recall	F1-Score
Claude-3.5-Sonnet	0.507	0.966	0.665
GPT-4o	0.437	0.941	0.597
Llama-3-70B	0.355	0.966	0.519

Tabla 3 Accuracy metrics for title and abstract-based screening.

	Precision		F1-Score
Claude-3.5-Sonnet	0.639	0.898	0.746
GPT-4o	0.6581	0.8644	0.7473
Llama-3-70B	0.413	0.890	0.565

The results show that Claude-3.5-Sonnet excelled with precision (0.507), F1-Score (0.665), recall (0,966), and 0.38% FN in the title-based classification process, while GPT-4o showed better balance with precision (0.6581), F1-Score (0.7473), recall (0,8644), and 1.52% FN when including the abstract.

Discussion

Evaluation of LLMs in the screening process reveals promising results. The high recall of all models (≥0.8644), with Claude-3.5-Sonnet standing out by including 96.6% of relevant articles, suggests their feasibility in initial screening automation. However, the low precision (≥0.355) indicates a tendency to include non-relevant articles, underscoring the need for human intervention at later stages. Despite this limitation, LLMs could significantly reduce the volume of articles requiring manual review, optimizing process time. F1-Score values, particularly for Claude-3.5-Sonnet and GPT-4o (≥0.74), are close to the recommended threshold of 0.8 (Lipton et al., 2014), suggesting potential for future improvements. Overall, these findings support using LLMs as valuable tools in optimizing screening in systematic literature studies, pointing to the importance of a hybrid approach that combines automation and human judgment.

Conclusions

In this study, we examined the efficacy of LLM models in automating the screening process for systematic literature reviews. Although the assessed LLMs exhibited high performance in identifying relevant articles, the presence of false negatives indicates that complete reliance on LLMs for screening is not yet feasible. Given their high sensitivity and the primary goal of excluding non-relevant articles, we propose that LLMs can effectively conduct the initial screening phase based on article titles.

Limitations and future work

This study found limitations mainly related to the number of requests and tokens supported by each model's APIs. Future work could evaluate the effectiveness of the current models in the automatic screening process using other promising prompting techniques such Few-Shot (Brown et al., 2020) and Chain-of-Thought (Wei et al., 2023).

Acknowledgments

This article received support from projects H2O Learn (PID2020-112584RB-C31) and GENIELearn (PID2023-146692OB-C31), funded by MCIN/AEI/10.13039/501100011033/

European Union.

References

Anthropic. (2024, June 20). Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan,

A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Castillo-Segura, P., Alario-Hoyos, C., Kloos, C. D., & Fernández Panadero, C. (2023). Leveraging the Potential of Generative AI to Accelerate Systematic Literature Reviews: An Example in the Area of Educational Technology. 2023 World

Engineering Education Forum - Global Engineering Deans Council (WEEFGEDC), 1–8. https://doi.org/10.1109/WEEF-GEDC59520.2023.10344098

Hugging Face. (n.d.). MMLU Pro—A Hugging Face Space by TIGER-Lab. Retrieved July 9, 2024, from https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-AbstractConference.html

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.01268

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding Classifiers to Maximize F1 Score. arXiv: Machine Learning. https://www.semanticscholar.org/paper/Thresholding-Classifiers-to-Maximize-F1-Score-LiptonElkan/0fc904dbde45f9e1b696c34b389b6e880094379d

Medeiros, R. P., Ramalho, G. L., & Falcão, T. P. (2019). A Systematic Literature Review on Teaching and Learning Introductory Programming in Higher Education. IEEE Transactions on Education, 62(2), 77–90. IEEE Transactions on Education. https://doi.org/10.1109/TE.2018.2864133

Meta. (2024, April 18). Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/

Muñoz, M., Cruz, L., Herrera, E., Jiménez, J., Muñoz, A., & Ramos, D. (2020). Pensamiento Computacional para la formación de maestros: Una revisión sistemática de literatura. Proceedings of the 18th LACCEI International Multi-Conference for Engineering, Education, and Technology: Engineering, Integration, And Alliances for A Sustainable Development" "Hemispheric Cooperation for Competitiveness and Prosperity on A Knowledge-Based Economy." The 18th LACCEI International Multi-Conference for Engineering, Education, and Technology: Engineering, Integration, And Alliances for A Sustainable Development" "Hemispheric Cooperation for Competitiveness and Prosperity on A Knowledge-Based Economy." https://doi.org/10.18687/LACCEI2020.1.1.135

Oliveira, L., Rosa, S. S., & Pimentel, A. (2019). Revisão Sistemática da Literatura: Formação de Grupos na Aprendizagem Colaborativa com Suporte Computacional. Anais Do XXX Simpósio Brasileiro de Informática Na Educação (SBIE 2019), 1955. https://doi.org/10.5753/cbie.sbie.2019.1955

OpenAI. (2024, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007 Sachs, N. A. (2018). Here's Some Great Research! Now What? Translating Research Into Practice. HERD: Health Environments Research & Design Journal, 11(1), 40–42. https://doi.org/10.1177/1937586718757309

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2024). The Prompt Report: A Systematic Survey of Prompting Techniques (arXiv:2406.06608). arXiv. http://arxiv.org/abs/2406.06608

Shin, T., Razeghi, Y., Logan Iv, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4222–4235. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.346

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903