II Congreso Latinoamericano y del Caribe de Innovación en Investigación en Educación Superior LatinSoTL, Universidad de Cuenca, Ecuador, y Universidad Peruana Cayetano Heredia, Lima, Perú, del 30 de septiembre al 4 de octubre de 2024.

¿Pueden los Grandes Modelos de Lenguaje Automatizar los Estudios Sistemáticos de Literatura? Explorando la Inclusión y Exclusión Automatizada – Un estudio de caso en el ámbito de la enseñanza de Ciencias de la Computación

Franklin L. Sánchez[1,2] [0000-0003-3963-2630], Carlos Alario-Hoyos[1] [0000-0002-3082-0814], Danny S. Guamán[2] [0000-0003-2794-3079], Julio C. Caiza[2] [0000-0001-9910-582X]

[1] Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Leganés, España.
[2] Escuela Politécnica Nacional, Av. Ladrón de Guevara E11-253, Quito, Ecuador
franklin.sanchez@alumnos.uc3m.es

Resumen. Este estudio evalúa la eficacia de los Grandes Modelos de Lenguaje en el proceso de inclusión/exclusión de Estudios Sistemáticos de Literatura en el área de la enseñanza de Ciencias de la Computación, dominio que cada vez presenta un mayor incremento en el número de contribuciones. Utilizando modelos como GPT-4o, Claude-3.5-Sonnet y Llama-3-70B se explora la automatización del proceso de inclusión/exclusión de artículos, comparando sus resultados con un proceso manual llevado a cabo por investigadores en el área. Los datos con los que se trabajó son de julio de 2024 y los resultados del proceso de selección muestran alta sensibilidad (≥0.8644) en todos los modelos, indicando que se incluyen al menos el 86% de los artículos relevantes, y se destaca que Claude-3.5-Sonnet incluye el 96,6% de los artículos relevantes. Los valores F1-Score para Claude-3.5-Sonnet y GPT-4o (≥0.74) muestran que el rendimiento de los modelos es aceptable para el contexto de este estudio. Aunque la baja precisión (≥0.355) indica que los modelos tienden a incluir artículos no relevantes, los resultados obtenidos sugieren que los LLM tienen potencial significativo como herramientas de apoyo en la etapa de inclusión/exclusión, pudiendo reducir el tiempo de revisión manual. No obstante, se recomienda un enfoque híbrido que combine automatización con juicio humano en las tareas finales de dicha etapa.

Palabras clave: Enseñanza de Ciencias de la Computación, Grandes Modelos de Lenguaje, Revisión Sistemática de literatura.

1 Introducción

El cuerpo de conocimiento (BoK) en la enseñanza de Ciencias de la Computación (CC) ha crecido exponencialmente, evidenciado por la proliferación de publicaciones en áreas como como el pensamiento computacional [1], estrategias pedagógicas en programación [2] y la formación de grupos de aprendizaje asistidos por computadora [3].

La aplicación de la inteligencia artificial en educación ha acelerado aún más este crecimiento, generando un flujo continuo de innovaciones [4].

Los Estudios Sistemáticos de Literatura (ESL) son cruciales para organizar este vasto BoK, determinar la madurez de las contribuciones y sintetizar tendencias emergentes. Sin embargo, la elaboración de un ESL requiere típicamente entre 12 y 18 meses [5], retrasando la aplicación de nuevos conocimientos en escenarios educativos reales y creando una brecha entre la generación de conocimiento y su implementación práctica.

Los Modelos de Lenguaje de Gran Escala (LLM) han surgido como herramientas prometedoras para automatizar fases de los ESL. Investigaciones previas como la de Castillo et all. en [6] han explorado su uso en Revisiones Sistemáticas de Literatura, con resultados alentadores pero mejorables.

Nuestra propuesta se distingue por utilizar modelos de vanguardia como GPT-4 [7], Claude-3.5-Sonnet [8] y Llama-3-70B [9], conocidos por su rendimiento superior a la fecha de la elaboración de este estudio [10], enfocándonos en automatizar el proceso de inclusión/exclusión de artículos en el contexto de un ESL en el área de la enseñanza de CC.

2 Marco teórico

Según Petersen et al. [11], los ESL consisten de tres fases principales: planificación, ejecución y reporte. En la fase de ejecución se definen las etapas de: Aplicar la estrategia de selección de artículos, aplicación de criterios de inclusión/exclusión, clasificación de artículos, análisis de artículos, y la identificación de lagunas y tendencias.

3 Método

Este estudio implementó las fases de selección de artículos e inclusión/exclusión de artículos de un ESL en el ámbito de la enseñanza de Ciencias de la Computación en educación superior, siguiendo las directrices propuestas en [11], como se ilustra en la Figura 1.

La fase de selección de artículos se realizó en marzo de 2024, utilizando la base de datos Scopus y se obtuvieron un total de 1055 artículos para su posterior evaluación.

La fase de inclusión/exclusión se ejecutó en dos etapas: evaluación basada en títulos, y análisis de títulos y resúmenes. Este proceso se llevó a cabo de manera paralela, tanto manualmente por tres investigadores expertos como de forma automatizada mediante los Modelos de Lenguaje de Gran Escala (LLM) seleccionados.

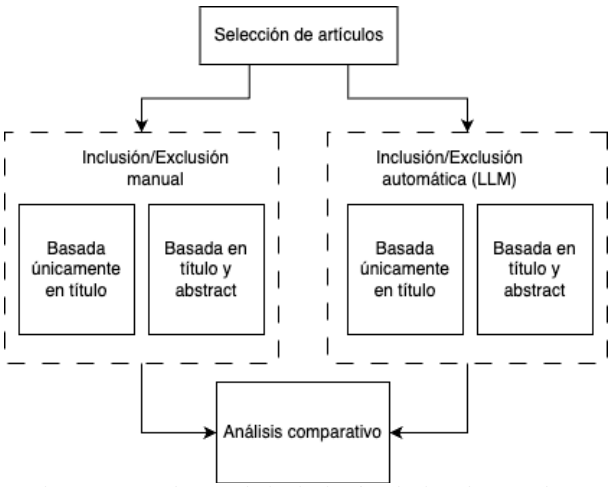

Figura 1 Método para la inclusión/exclusión de artículos

En la inclusión/exclusión manual, intervinieron 3 investigadores quienes aplicaron criterios de inclusión/exclusión predefinidos. Cada artículo fue revisado por uno de los investigadores, mientras que los otros 2 realizaron pilotos de prueba para garantizar la aplicación rigurosa de los criterios.

Para la inclusión/exclusión automatizada, se emplearon los tres LLM mejor rendimiento: GPT-4, Claude-3.5-Sonnet y Llama-3-70B. Se desarrollaron un scripts en Python que interactúan con las APIs de estos modelos, y se construyeron prompts basados en una plantilla [12] que incorpora algunos de los componentes comunes definidos en [13], aplicando la técnica de prompting zero-shot [14]. Para la inclusión/exclusión basada únicamente en el título se empleó el prompt definido en la Tabla 1, mientras que, para la inclusión/exclusión basada en el título y resumen, se creó y aplicó un prompt específico por cada criterio de inclusión establecido en el ESL, manteniendo una estructura consistente con el prompt inicial.

Tabla 1 Prompt para la inclusión/exclusión por título

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'.

Para evaluar la eficacia de los modelos en el proceso de screening, se construyeron matrices de confusión, a partir de las cuales se calcularon métricas de precisión, incluyendo sensibilidad, precisión y F1-Score. Estas métricas permitieron una comparación sistemática entre el desempeño de los LLM y el de los investigadores humanos, proporcionando una base cuantitativa para evaluar la viabilidad de la automatización en el proceso de screening.

4 Resultados

Los resultados para la clasificación en base al título se pueden observar en la Tabla 2, mientras que los resultados para la clasificación de los artículos en base al título y resumen se muestran en la Tabla 3.

Tabla 2 Métricas de precisión para la clasificación en base al título.

Modelo	Precisión	Sensibilidad	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 Métricas de precisión para la clasificación en base al título y resumen.

Modelo	Precisión	Sensibilidad	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

De los resultados se observa que Claude-3.5-Sonnet destacó con precisión (0.507), F1-Score (0.665), sensibilidad (0,966) y 0.38% de FN en el proceso de clasificación basado en el título, mientras GPT-4o mostró mejor equilibrio con precisión (0.6581), F1-Score (0.7473), sensibilidad (0,8644) y 1.52% de FN al incluir el resumen.

5 Discusión

La evaluación de los LLM en el proceso de inclusión/exclusión revela resultados prometedores. La alta sensibilidad (≥0.8644) de todos los modelos, con Claude-3.5-Sonnet destacando al incluir el 96,6% de los artículos relevantes, sugiere su viabilidad en la automatización inicial del screening. Sin embargo, la baja precisión (≥0.355) indica una tendencia a incluir artículos no relevantes, subrayando la necesidad de intervención humana en fases posteriores. A pesar de esta limitación, los LLM podrían reducir significativamente el volumen de artículos que requieren revisión manual, optimizando el tiempo del proceso. Los valores de F1-Score, particularmente para Claude-3.5-Sonnet

y GPT-4o (≥0.74), se acercan al umbral recomendado de 0.8 [15], sugiriendo potencial para mejoras futuras. En conjunto, estos hallazgos respaldan el uso de LLM como herramientas valiosas en la optimización del screening en Estudios Sistemáticos de Literatura, señalando la importancia de un enfoque híbrido que combine automatización y juicio humano.

6 Conclusiones

En este trabajo se evaluó la eficacia de modelos de LLM en el proceso automático de inclusión/exclusión de un estudio sistemático de literatura, que aunque han demostrado tener un alto nivel de desempeño para la tarea de inclusión/exclusión de artículos, la existencia de falsos negativos hace que el proceso de screening todavía no se pueda confiar enteramente en ellos. No obstante, dada la alta sensibilidad de los modelos y que la principal tarea de un proceso de screening es excluir artículos no relevantes, se estima que actualmente los LLM pueden ejecutar de manera automática, una primera fase de inclusión/exclusión en base al título.

7 Limitaciones y Futuras Investigaciones

Durante este estudio se encontraron limitaciones principalmente relacionadas con el número de peticiones y tokens soportados por las APIs de cada modelo. En trabajos futuros se podría evaluar la eficacia de los modelos actuales en el proceso de screening automático usando otras técnicas prometedoras de prompting como Few-Shot [16] y Chain-of-Thought [17].

Agradecimientos

Este trabajo ha sido apoyado por los proyectos H2OLearn (PID2020-112584RB-C31) y GENIELearn (PID2023-146692OB-C31), financiados por MCIN/AEI/10.13039/501100011033/Unión Europea.

Referencias

1. 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
2. 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
3. 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
4. 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
5. 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
6. 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 (WEEF-GEDC)*, 1–8. https://doi.org/10.1109/WEEF-GEDC59520.2023.10344098
7. *Hello GPT-4o*. (n.d.). Retrieved July 9, 2024, from https://openai.com/index/hello-gpt-4o/
8. *Introducing Claude 3.5 Sonnet*. (n.d.). Retrieved July 9, 2024, from https://www.anthropic.com/news/claude-3-5-sonnet
9. *Introducing Meta Llama 3: The most capable openly available LLM to date*. (n.d.). Retrieved July 9, 2024, from https://ai.meta.com/blog/meta-llama-3/
10. *MMLU Pro—A Hugging Face Space by TIGER-Lab*. (n.d.). Retrieved July 9, 2024, from https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro
11. 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
12. 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
13. 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

14. 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-Abstract-Conference.html
15. 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-Lipton-Elkan/0fc904dbde45f9e1b696c34b389b6e880094379d
16. 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
17. 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