Zero-Shot RCT Identification using Large Language Models: A Comparative Study with the Cochrane Classifier

SA107

Seye Abogunrin[1], **Marie Lane**[1], **Roberto Rey Sieiro**[2]

1 F. Hoffmann-La Roche Ltd, Basel, Switzerland
2 Roche Farma, S.A., Madrid, Spain

Takeaway

A zero-shot, prompt-engineered large language model (LLM) identifies randomised controlled trials (RCTs) from titles and abstracts with greater overall accuracy and balance than the specialised Cochrane classifier.

Background

- Identifying relevant study types is a critical and time-consuming step in conducting systematic reviews. [1,2,13]
- Automating RCT identification can accelerate the review process by improving record filtering and workload assignment for review teams. [3,4]
- Specialised tools, such as the Cochrane classifier, have been the traditional approach for this task, relying on machine learning models trained on specific datasets. [3,5]
- Modern, general-purpose LLMs like GPT-4.1 offer a powerful and flexible alternative, capable of performing complex classification tasks with no prior task-specific training (i.e., "zero-shot"). [6–8]
- This study compares a general-purpose LLM against a specialised benchmark tool to assess its effectiveness for RCT identification in a systematic review workflow.

Methods

- A dataset of 2380 titles and abstracts, TIABs, manually screened for a breast cancer systematic literature review, was used for this study.
- A specialised prompt was developed to instruct a GPT-4.1 model to classify records as either a likely RCT or not likely an RCT in a zero-shot workflow.
- The pipeline initialises an LLM model (GPT-4.1) with adjustable parameters such as temperature (which influences the balance between predictability and creativity in generated text) to manage model behaviour. [10]
- The temperature parameter was set to 0.1 to standardise model behaviour.
- The same dataset was also processed using the established Cochrane classifier for a direct comparison.
- Key performance metrics—accuracy, precision, recall, and F1-score—were calculated to provide a comprehensive assessment of each tool's classification capabilities. [9]
- The performance of both classifiers was evaluated against a human-labeled ground truth.

Results

- The prompt-engineered LLM classifier demonstrated superior overall performance versus the ground truth, achieving an F1-score of 0.74, considerably outperforming the Cochrane classifier's score of 0.49.
- The LLM classifier showed a balanced performance profile with high accuracy (91.1%), high precision (78%), and strong recall (70%).
- The Cochrane classifier achieved a higher recall (86%) but at the cost of very low precision (33.7%) and overall accuracy (67.4%).
- The Cochrane classifier's low precision resulted in a high volume of false positives, incorrectly identifying a large number of non-RCTs as RCTs. [3]

Classifier	Accuracy	Recall	Precision	F1-score
Zero-Shot LLM	0.91	0.70	0.78	0.74
Cochrane	0.67	0.86	0.34	0.49

Table 1: Results for the breast cancer data set, 2380 abstracts

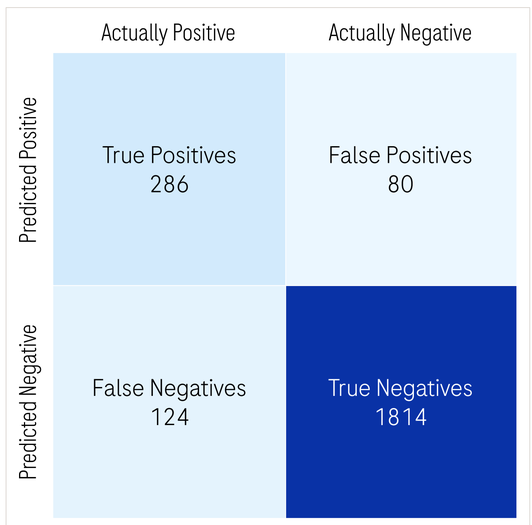

Figure 1: Confusion Matrix for Zero-Shot LLM Classification

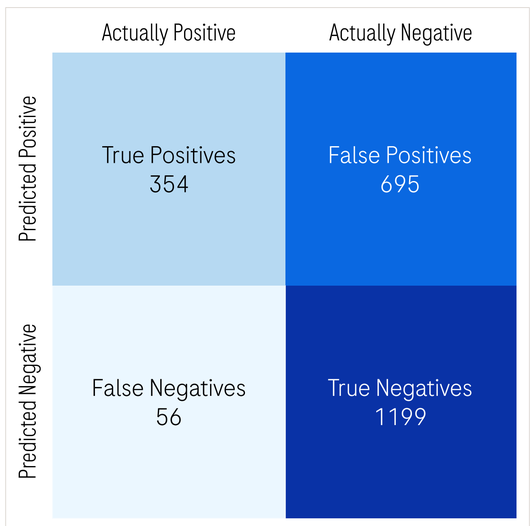

Figure 2: Confusion matrix for Cochrane Classification

Discussion

- The study confirmed that a general-purpose LLM classifier guided by a well-designed prompt can be a more balanced and robust solution for RCT identification than a specialised tool. [3,11,12]
- The Cochrane classifier's optimisation for high recall leads to a significant number of false positives, which can increase the workload for human reviewers in a double-screening scenario.
- The LLM classifier's superior F1-score highlights its ability to provide a more reliable and better balance between recall and precision for identifying likely RCTs, enabling review teams to manage datasets more efficiently.
- The LLM classifier's strong performance was achieved with a zero-shot approach, requiring only a dedicated prompt. This demonstrates that review teams can develop their own classifiers without relying on the Cochrane classifier, while still avoiding the need for model retraining.
- Reproducibility concerns were mitigated by lowering the temperature parameter, which reduced the randomness in the LLM's outputs. [10]
- Incorporating the LLM classifier into a literature review management system allows ease of data connectivity and minimal disruption to the screening workflow
- We propose that classifying titles and abstracts as likely-RCT or not-RCT, could be utilised to support screening, for example, not-RCTs could be quickly excluded with an acceptable low risk of missing eligible RCTs in suitable reviews (i.e. reviews where non RCT study designs are not eligible).
- A limitation of this study is that it was conducted using one dataset, further research on additional datasets could assess if these findings are consistent for other reviews.

Conclusion

- A prompt-engineered GPT-4.1 provides a more balanced and effective solution for RCT identification than the Cochrane classifier.
- The LLM classifier's substantially higher F1-score demonstrates its ability to effectively distinguish RCTs while minimising the false positives that can burden review teams.
- The LLM classifier's superior overall performance, achieved with a dedicated prompt without model training, makes it a more practical and efficient tool for resource-constrained systematic review workflows.

References:

1. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Med. 2010;7(9):e1000326. doi:10.1371/journal.pmed.1000326.
2. Shojania KG, Sampson M, Ansari MT, et al. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. Ann Intern Med. 2007;147(4):224-233. doi:10.7326/0003-4819-147-4-200708210-00179.
3. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021;133:140-151. doi:10.1016/j.jclinepi.2020.11.003.
4. Cochrane Handbook – Chapter 22. Prospective approaches to accumulating and maintaining evidence. Version 6.5 (2024).
5. EPPI-Reviewer Team. Machine learning functionality in EPPI-Reviewer (v10, web version). UCL; 2025.
6. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. NeurIPS 2020. arXiv:2005.14165.
7. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. NeurIPS 2022. arXiv:2205.11916.
8. Li M, Sun J, Tan X, et al. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Systematic Reviews. 2024;13(1):219. doi:10.1186/s13643-024-02609-x.
9. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008.
10. OpenAI. Advanced usage: controllability, temperature and sampling. OpenAI Platform Docs. Link: https://platform.openai.com/docs/guides/advanced-usage/controlling-responses
11. Cochrane MECIR (Methods): Selecting studies to include in the review (C39–C42).
12. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019;19:132. doi:10.1186/s12874-019-0782-0.
13. Abogunrin S, Muir JM, Zerbini C, Sarri G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. Front Pharmacol. 2025;16:1454245. doi:10.3389/fphar.2025.1454245.