Calibrating Large Language Model Probabilities to Achieve Target Recall in Systematic Review Screening

SA18

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 calibration workflow can transform raw large language model (LLM) probability outputs into a reliable, tunable classifier, enabling users to achieve specific recall targets for systematic literature review (SLR) screening.

Background

- The ever-increasing volume of published medical research makes the manual identification of specific study designs, like randomised controlled trials (RCTs), for SLRs a significant bottleneck in generating timely clinical evidence. [1]
- While generative AI (GenAI) shows promise for automating the title and abstract (TIAB) screening phase, standard LLM classifiers lack the necessary control for high-stakes applications where methodological rigor is paramount. [2,6]
- In SLRs of interventions, missing an eligible record (a failure of recall) is often considered a more critical error than incorrectly including a non-eligible record (a failure of precision), as it can bias the final treatment effect estimates. [3,4]
- Standard LLM outputs are often overconfident and do not provide a mechanism to prioritize recall. This limitation hinders their adoption in evidence synthesis, where the goal is comprehensive evidence retrieval. [5,7–9]
- This experiment evaluates a workflow designed to calibrate an LLM's (GPT-4o) probabilistic outputs, creating a "control knob" to tune its performance specifically for the task of identifying eligible RCT TIAB records to a pre-defined recall target. [8,9]

Methods

- A human-labelled dataset of 2380 TIABs screened from a previous SLR was utilised. The primary objective of the screening exercise was to identify all relevant RCTs based on the SLR protocol. Each record was labelled "Include" (is a relevant RCT) or "Exclude.".
- The dataset was randomly split into three sets:
  - Training Set (20%): Used to train the calibration model.
  - Validation Set (20%): Used to select the optimal decision threshold for identifying RCTs.
  - Test Set (60%): Withheld for the final, unbiased performance evaluation.
- A multi-stage workflow was implemented:
  1. Probability Extraction: The LLM (GPT-4o) was prompted to classify if a record was an RCT and to generate token-level log probabilities (log (P)) for its "Include" and "Exclude" decisions. These were then converted to class probabilities. [10–11]
  2. Calibration Model Training: An isotonic regression model was fitted on the training set to map the LLM's raw, often miscalibrated, probabilities to more reliable scores. [8,9]
  3. Threshold Selection: The calibrated model was applied to the validation set. Performance for RCT identification was assessed across all possible decision thresholds to find the optimal point that met our desired recall target (as visualized in the provided plot).
  4. Final Evaluation: The final calibrated model and the chosen threshold were applied to the unseen test set to measure the final recall and precision for identifying likely RCTs.

Results

- The calibration workflow successfully corrected the LLM's raw probability outputs, transforming them into a reliable probability score for identifying RCTs.
- A specific decision threshold was chosen from the validation set performance to maximize the identification of RCTs (high recall), as shown in Figure 3.
- The reliability of the final workflow was validated on the withheld test set, with results summarised in Table 1. The model achieved the following performance metrics:
  - Recall: 83% (correctly identifying 83% of all true RCTs)
  - Precision: 56% (of all studies flagged as RCTs, 56% were correct)
- These results confirm that the model can be configured to operate at a high-recall threshold, which is crucial for minimising the risk of excluding relevant RCTs from the review. The trade-off is a moderate precision score, ensuring a conservative approach.
- Using the 0.1 probability threshold in our example shows that the reviewer would have had to screen 162 extra full-texts (and less studies excluded wrongly; Figure 1). This aligns more with practices of reviewers where records are included when the reviewer is unclear about whether the record should be included or excluded.
- In comparison, the default 0.5 probability threshold (i.e., without calibration or threshold tuning) produced a substantially lower recall, as shown in Table 1, indicating that this configuration is unsuitable for recall-sensitive tasks such as identifying RCTs in SLRs. However, this would have led to screening of an extra 21 full-texts (Figure 2).

Threshold	Model	Recall	Precision	F1-score
0.1	GPT4o	0.83	0.56	0.67
0.5 (Default)	GPT4o	0.70	0.89	0.79

Table 1: Results for the test data set, 1428 abstracts

	Actually Positive	Actually Negative
Predicted Positive	True Positives 204	False Positives 162
Predicted Negative	False Negatives 42	True Negatives 975

Figure 1: Confusion matrix with 0.1 threshold

	Actually Positive	Actually Negative
Predicted Positive	True Positives 173	False Positives 21
Predicted Negative	False Negatives 73	True Negatives 1116

Figure 2: Confusion matrix with 0.5 (Default) threshold

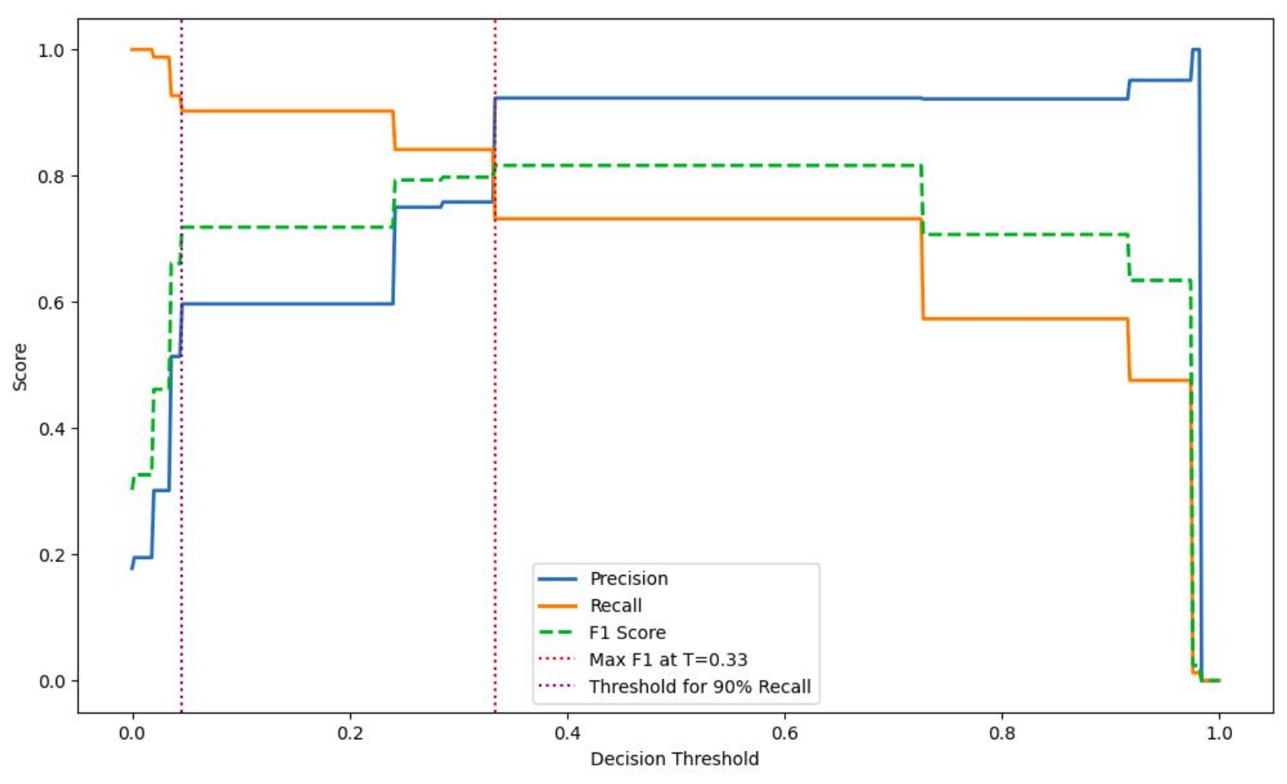

Figure 3: Effect of threshold on performance for validation set

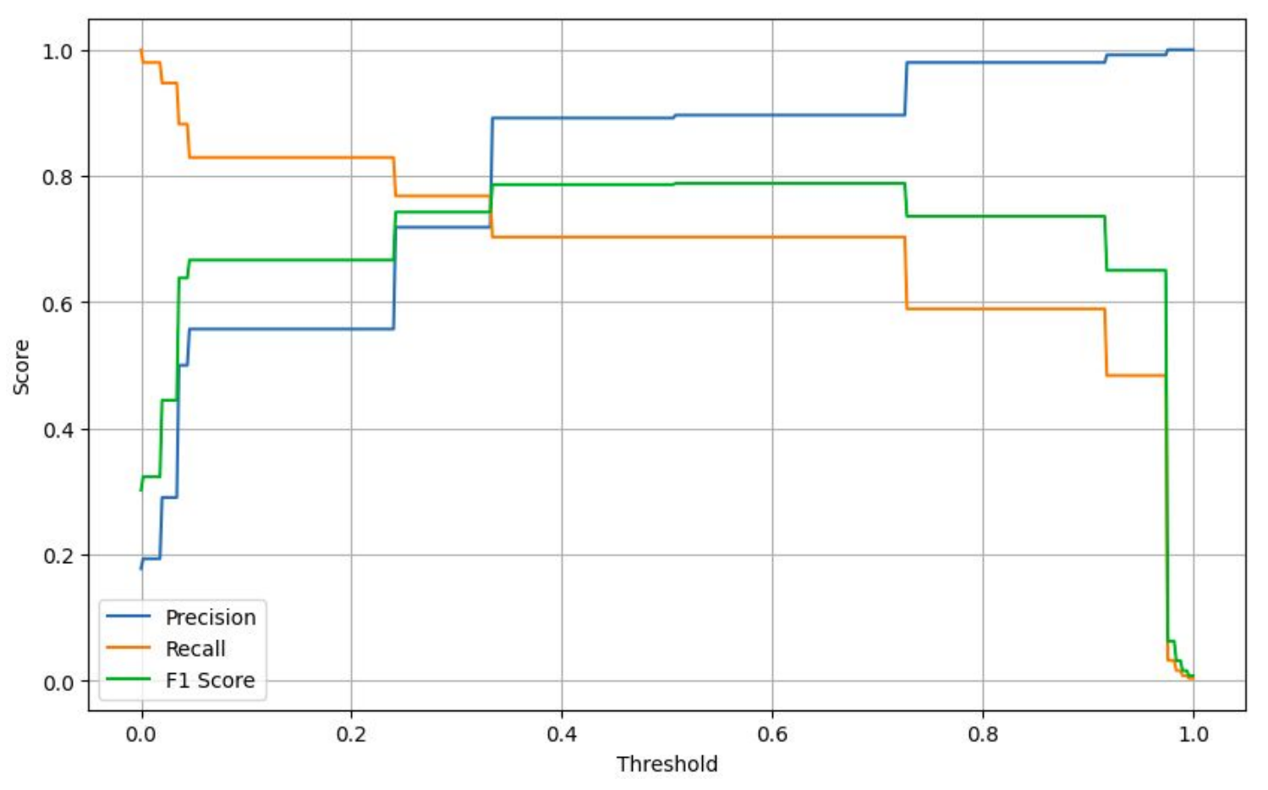

Figure 4: Effect of threshold on performance for test set

Discussion

- This experiment demonstrates a critical advancement for AI use in evidence synthesis: moving from simple classification to controlled, reliable, and tunable screening, specifically for the "gold standard" evidence of SLRs. [4,5]
- The ability to set a high recall target directly addresses a primary concern of researchers and regulatory bodies regarding the use of AI in SLRs—the fear of missing pivotal evidence.
- Crucially, the performance achieved through calibration and threshold tuning significantly surpasses the unoptimised results from a default 0.5 probability cutoff. This highlights the necessity of such a workflow for practical deployment in SLRs.
- An 83% recall for RCTs ensures the vast majority of relevant trials proceed to full-text review. The corresponding 56% precision still filters out a significant portion of non-RCTs, drastically reducing the manual workload compared to a fully manual screen.
- The calibrated workflow could be applied to the same eligibility criteria for other reviews.
- This "control knob" is especially useful for managing a workflow involving a single reviewer assisted with an AI-reviewer, as it serves as a safeguard to prevent eligible publications from being mistakenly classified as ineligible.
- The workflow was optimised to maximise recall, which does not necessarily minimise reviewer workload; [12] focusing solely on workload reduction or a single performance metric (e.g., F1 score) may overlook other important factors such as tool sensitivity, user proficiency, and practical trade-offs.

Conclusion

- Raw LLM outputs are not reliable for automated classification but a robust calibration workflow resolves this issue.
- By calibrating probabilities and selecting a strategic decision threshold, an LLM can be transformed into a specialised screening tool tuned to achieve high recall for specific eligibility criteria, which can be used across many SLRs.
- This methodology provides the control and transparency needed to confidently integrate the calibrated outputs from LLM processing into the SLR workflow, ensuring both efficiency and methodological rigor.
- This approach strengthens the case for using GenAI to accelerate evidence generation, allowing researchers to focus on higher-level synthesis and analysis of eligible evidence to inform clinical and policy decisions.

References:

1. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Medicine. 2010;7(9):e1000326. doi:10.1371/journal.pmed.1000326.
2. 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.
3. Cochrane Handbook for Systematic Reviews of Interventions, v6.5 (2024), Chapter 13. Assessing risk of bias due to missing evidence.
4. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
5. Cochrane Handbook for Systematic Reviews of Interventions, v6.5 (2024), Chapter 22. Prospective approaches to accumulating and maintaining evidence.
6. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Proceedings of ICML 2017. PMLR; 2017:1321–1330.
7. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.
8. Niculescu-Mizil A, Caruana R. Predicting Good Probabilities with Supervised Learning. In: ICML '05. ACM; 2005:625–632. doi:10.1145/1102351.1102430.
9. Zadrozny B, Elkan C. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In: KDD '02. ACM; 2002:694–699. doi:10.1145/775047.775151.
10. OpenAI Cookbook. Using logprobs with Chat Completions. Link: https://platform.openai.com/docs/guides/advanced-usage/controlling-responses
11. Microsoft Learn (Azure OpenAI). logprobs parameter for Chat Completions.
12. Abogunrin S, Slob BPH, Lane M, Emamipour S, Twardowski P, Boersma C, van der Schans J. The introduction and adoption of artificial intelligence in systematic literature reviews: a discrete choice experiment. BMJ Open. 2025;15(10):e099921. doi:10.1136/bmjopen-2025-099921.