# Generative Artificial Intelligence: An Effective Alternative for Screening Titles and Abstracts in Systematic Literature Reviews

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

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

MSR224

## Background

The increase in published research articles makes conducting Systematic Literature Reviews (SLRs) more challenging and time-consuming.[1]

Title and abstract (TIAB) screening, a crucial step in SLRs, often becomes a bottleneck due to the large number of studies requiring careful manual review.[2]

Traditional machine learning algorithms for automating TIAB screening need extensive labeled data and substantial computational resources, limiting their practicality, especially in fields where such data is scarce.[3]

Generative Artificial Intelligence (GenAI) models (e.g., OpenAI's GPT, Google's Gemini, Meta's LLAMA, and Anthropic's Claude) offer an alternative by using advanced language understanding to automate TIAB screening with minimal labeled data.[4]

Evaluating different GenAI models and prompting methods is essential to find the most effective approach for automating TIAB screening, potentially transforming SLRs.[5]

Automating TIAB screening with GenAI reduces researchers' workloads and accelerates evidence generation, facilitating quicker market access for new medicines and ultimately benefiting patient care.

This experiment focuses on using GPT models (i.e., GPT-3.5, GPT-4 Turbo, and GPT-4o) to accelerate the TIAB screening step in the SLR process.

## Methods

Human-labeled data from the TIAB screening step were obtained from three SLRs covering distinct therapeutic areas: non-small cell lung cancer (NSCLC), castration-resistant prostate cancer (CRPC), and COVID-19.

- The SLR protocol defined the inclusion and exclusion criteria using the Population, Intervention, Comparator, Outcomes, Study Design (PICOS) framework.
- Each record was labeled with an inclusion or exclusion status + exclusion reason.

A subset of 50 examples, randomly sampled, from each SLR was utilized to fine-tune the prompts used in the analysis. These examples were not part of the test set.

Two main methods were evaluated for automating the TIAB screening, see table 1

The same pipeline architecture was used for each of the SLR, although prompts differed. All of them had the same categories for exclusion criteria; PICOS.

Based on the provided information, Method1 and Method 2 generates a decision to "Include" or "Exclude" each study and offers an "Exclusion Reason" for each criterion, if applicable, along with a confidence score.

The results from both methods were compared at the binary inclusion and exclusion level versus the human labeled dataset.

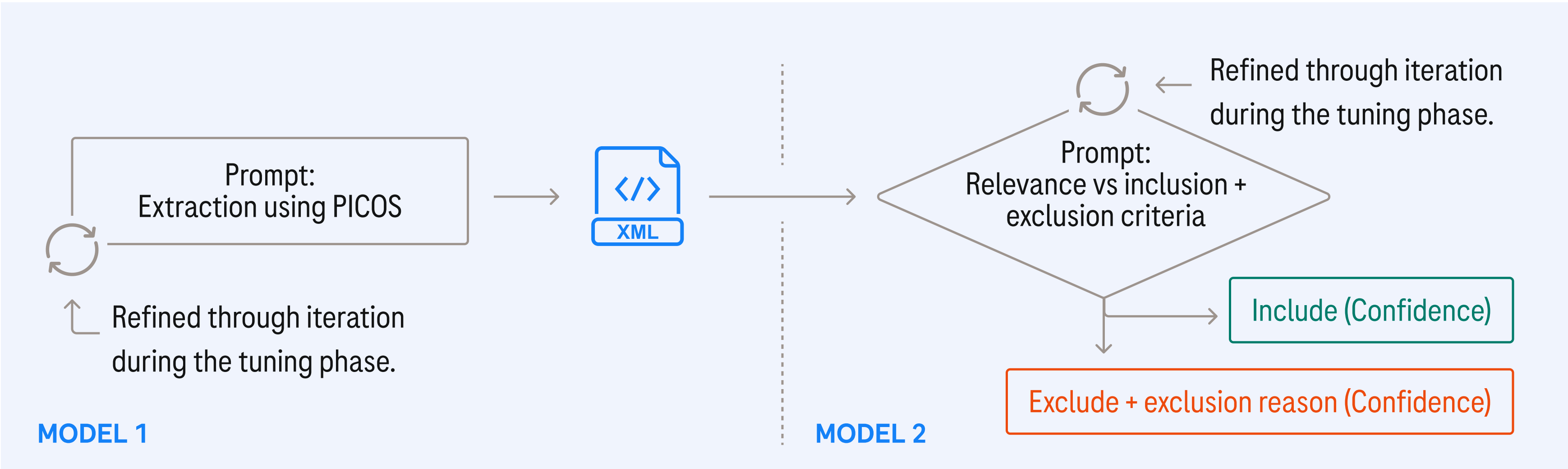
Both pipelines start by initializing various Azure-based Large Language Model (LLM) models (i.e., GPT-3.5, GPT-4 Turbo, GPT-4o) with adjustable parameters such as temperature (a parameter influencing the balance between predictability and creativity in generated text) to manage model behavior.

### Method 1

Uses prompts to extract relevant data, followed by additional prompting to include or exclude studies based on the extracted information.

Using prompt engineering techniques, these LLM models extract relevant information from the abstracts, organizing it into XML based on the PICOS criteria.

A second LLM model is then initialized to process the extracted information. The data is provided in XML format, alongside specific questions aligned with the PICOS criteria of the respective SLR.



Prompt: Extraction using PICOS

Refined through iteration during the tuning phase.

XML

Refined through iteration during the tuning phase.

Prompt: Relevance vs inclusion + exclusion criteria

Include (Confidence)

Exclude + exclusion reason (Confidence)

MODEL 1

MODEL 2

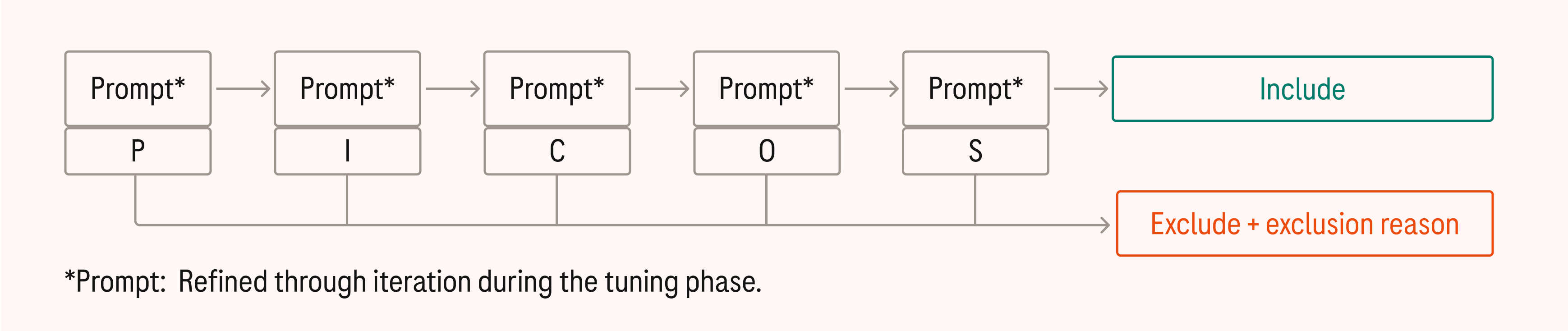
### Method 2

Tests hierarchical prompting with two sub-approaches:

- Complex prompting, which incorporated all inclusion and exclusion criteria.
- Simpler prompting using Natural Language Constructed Prompts (NLCP), focusing only on inclusion criteria.

Using prompt engineering technique, this method evaluates each category of the PICOS criteria separately.

For each category, a prompt is provided to assess inclusion. If any criterion leads to exclusion, the final decision is marked as "Exclude." The "Exclusion Reason" is then determined by applying a hierarchical structure specific to the SLR.



Prompt* P

Prompt* I

Prompt* C

Prompt* O

Prompt* S

Include

Exclude + exclusion reason

*Prompt: Refined through iteration during the tuning phase.

Sub-approach 1 provides the model with all inclusion and exclusion criteria in a structured, bullet-point format within the prompt.

Sub-approach 2 reformulates the inclusion criteria into natural language questions, allowing the model to evaluate each criterion more intuitively.

Figure 1: Overview of the pipeline for Method 1 and Method 2

## Results

The SLR datasets evaluated in this study were previously utilized to address research questions in NSCLC, CRPC, and COVID-19.

The pipeline that yielded the best results was method 2 that used natural language constructed prompts (NLCP) that considered the inclusion reasons alone.

- The pipelines were evaluated with "Conflict Rate", which is the disagreement rate between the AI model's decisions and the human labels, also calculated as 100 - % Accuracy. Due to the fact that SLR datasets are imbalanced, with a huge proportion of excludes, additional metrics were also calculated to provide a full picture of the models' performance.
- The conflict rate for this pipeline was 6%, 12% and 14% for each of the SLR respectively.
- Using GPT-3.5 yielded an average of 10% more conflicts but there were not big differences between GPT- 4 Turbo and GPT- 4o in terms of conflict rate. Although GPT- 4o offered a significant advantage in terms of speed and cost, being twice as fast and more economical compared to GPT-4 Turbo.

GPT-4o outperformed GPT-4 Turbo in terms of speed (GPT-4o processed the dataset twice as fast) and cost-effectiveness[6] (at the time of the experiment the cost per 1 million input tokens was $5 vs $10), without sacrificing accuracy in this particular experiment.

**Takeaway**

Generative AI models are effective for screening titles and abstracts in systematic literature reviews (SLRs).

### NSCLC

| Method | Model | Recall | Precision | Specificity | Accuracy | Conflicts |
|---|---|---|---|---|---|---|
| Method 1 - Extraction and Prompting | GPT-3.5 | 0.48 | 0.5 | 0.95 | 0.90 | 10% |
| | GPT-4 Turbo | 0.46 | 0.55 | 0.94 | 0.91 | 9% |
| | GPT-4o | 0.48 | 0.6 | 0.96 | 0.91 | 9% |
| Method 2 - Complex Prompting | GPT-3.5 | 0.4 | 0.2 | 0.81 | 0.77 | 23% |
| | GPT-4 Turbo | 0.3 | 0.65 | 0.98 | 0.91 | 9% |
| | GPT-4o | 0.18 | 0.49 | 0.98 | 0.9 | 10% |
| Method 2 - NLCP | GPT-3.5 | 0.04 | 0.36 | 0.99 | 0.89 | 11% |
| | GPT-4 Turbo | 0.75 | 0.72 | 0.97 | 0.94 | 6% |
| | GPT-4o | 0.8 | 0.66 | 0.95 | 0.94 | 6% |

Table 1: Results for the NSCLC data set, 8773 abstracts, test dataset 8723

### CRPC

| Method | Model | Recall | Precision | Specificity | Accuracy | Conflicts |
|---|---|---|---|---|---|---|
| Method 1 - Extraction and Prompting | GPT-3.5 | 0.9 | 0.49 | 0.68 | 0.74 | 26% |
| | GPT-4 Turbo | 0.47 | 0.79 | 0.9 | 0.79 | 21% |
| | GPT-4o | 0.84 | 0.63 | 0.83 | 0.84 | 16% |
| Method 2 - Complex Prompting | GPT-3.5 | 0.44 | 0.31 | 0.67 | 0.61 | 39% |
| | GPT-4 Turbo | 0.97 | 0.5 | 0.67 | 0.75 | 25% |
| | GPT-4o | 0.89 | 0.67 | 0.85 | 0.86 | 14% |
| Method 2 - NLCP | GPT-3.5 | 0.37 | 1 | 1 | 0.84 | 16% |
| | GPT-4 Turbo | 0.8 | 0.69 | 0.88 | 0.86 | 14% |
| | GPT-4o | 0.83 | 0.72 | 0.89 | 0.86 | 12% |

Table 2: Results for the CPRC data set, 3371 abstracts, test dataset 3321

### Covid 19

| Method | Model | Recall | Precision | Specificity | Accuracy | Conflicts |
|---|---|---|---|---|---|---|
| Method 1 - Extraction and Prompting | GPT-3.5 | 0.29 | 0.2 | 0.88 | 0.82 | 18% |
| | GPT-4 Turbo | 0.4 | 0.31 | 0.91 | 0.86 | 14% |
| | GPT-4o | 0.29 | 0.37 | 0.95 | 0.89 | 11% |
| Method 2 - Complex Prompting | GPT-3.5 | 0.21 | 0.18 | 0.9 | 0.84 | 16% |
| | GPT-4 Turbo | 0.3 | 0.25 | 0.91 | 0.85 | 15% |
| | GPT-4o | 0.26 | 0.22 | 0.91 | 0.84 | 16% |
| Method 2 - NLCP | GPT-3.5 | 0.04 | 0.28 | 0.99 | 0.9 | 10% |
| | GPT-4 Turbo | 0.24 | 0.29 | 0.94 | 0.87 | 13% |
| | GPT-4o | 0.3 | 0.27 | 0.92 | 0.86 | 14% |

Table 3: Results for the Covid 19 data set, 4968 abstracts, test dataset 4918

## Discussion

The study demonstrated the capability of GPT-3.5, GPT-4 Turbo, and GPT-4o in automating TIAB screening in SLRs required no effort to identify a validation sample and minimal effort to fine-tune the prompts.

Consistent with published research (beyond SLR) we see improvements as the models develop from GPT-3.5 to GPT-4o and GPT-4 Turbo. [7]

Although advanced prompt engineering techniques streamlined the TIAB screening process, the classification phase was completed in substantially less time compared to human-only.

Of the three prompting methods evaluated, NLCP method showed better performance, resulting in lower conflict rates across all datasets.

Reproducibility concerns were mitigated by the generation of transparent exclusion reasons which improved the robustness of the binary evaluation.

A limitation of this experiment is that it was conducted using retrospective SLR datasets, therefore future research should assess if these findings are consistent in a prospective SLR setting.

## Conclusion

Generative AI models, effectively automate the TIAB screening process in SLRs.

The NLCP method produced conflict rates of 6-14%, outperforming more complex prompting strategies.

Both GPT-4 Turbo and GPT-4o showed high accuracy, but GPT-4o was faster and more cost-efficient, making it the preferred choice.

As the NCLP prompts used natural language the use of the method could be adopted by a broader range of researchers than other prompting strategies, particularly if used via a user friendly interface in an SLR tool.

These GenAI approaches can accelerate the SLR process, enabling quicker evidence generation and facilitating market access for new medications.

References

1. J. C. Carver, E. Hassler, E. Hernandes and N. A. Kraft, "Identifying Barriers to the Systematic Literature Review Process," 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, Baltimore, MD, USA, 2013, pp. 203-212, doi: 10.1109/ESEM.2013.28.
2. Dennstädt, F., Zink, J., Putora, P.M. et al. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Syst Rev 13, 158 (2024). https://doi.org/10.1186/s13643-024-02575-4
3. Marshall, I.J., Wallace, B.C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 8, 163 (2019). https://doi.org/10.1186/s13643-019-1074-9
4. M. Mahbub Hossain "Using ChatGPT and other forms of generative AI in systematic reviews: Challenges and opportunities" Journal of Medical Imaging and Radiation Sciences, Volume 55, Issue 1, 11 - 12
5. Wang L, Bi W, Zhao S, Ma Y, Lv L, Meng C, Fu J, Lv H. Investigating the Impact of Prompt Engineering on the Performance of Large Language Models for Standardizing Obstetric Diagnosis Text: Comparative Study. JMIR Form Res. 2024 Feb 8;8:e53216. doi: 10.2196/53216. PMID: 38329787; PMCID: PMC10884897
6. https://openai.com/index/hello-gpt-4o/
7. https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu