Testing Automated Prompt Engineering Strategies for Systematic Literature Review Screening

SA92

Kim Wager, Gemma Carter, Christian Eichinger, Obaro Evuarherhe, Polly Field, Tomas Rees

Oxford PharmaGenesis, Oxford, UK

Scan here to download the poster

ISPOR main topic/taxonomy: Study Approaches
ISPOR subtopics: Artificial Intelligence, Machine Learning, Literature Review & Synthesis

INTRODUCTION

- Citation screening is one of the most laborious phases of systematic literature reviews, with researchers often evaluating hundreds or thousands of titles and abstracts.
- Large language models (LLMs) offer opportunities to reduce this burden.
- LLM performance varies significantly based on nuances in prompt design; strategic prompt-engineering techniques can significantly enhance LLM performance.

OBJECTIVE

- To assess how various prompt-engineering techniques influence LLM performance in citation-screening tasks versus basic instruction prompting (zero-shot prompting baseline).

METHODS

Dataset

- A previously published systematic review on high-resolution peripheral quantitative computed tomography (HR-pQCT) was used, which showed that HR-pQCT parameters can predict incident fracture.[1]
  - 534 titles and abstracts with human-screened ground truth labels.
- We excluded examples used in the few-shot prompts from the test set to ensure that the model was not evaluated on data it had already seen (n = 524).

Prompting techniques

- Prompting techniques tested are summarized in **Figure 1**; models tested were GPT-4o and Claude 3.5 Sonnet.

Evaluation metrics

- Inclusion recall: proportion of true includes correctly identified.
  - Critical for ensuring comprehensive evidence capture.
  - Low inclusion recall risks missing relevant studies and compromising review completeness.
- Exclusion recall: proportion of true excludes those that were correctly identified.
  - Essential for screening efficiency.
  - Low exclusion recall leads to wasted time reviewing irrelevant studies in subsequent review stages.
- Processing time (efficiency).

RESULTS

- Inclusion and exclusion recall across all models and strategies tests is shown in **Figure 2**.

Chain-of-thought (CoT) prompting

- Claude 3.5 Sonnet outperformed GPT-4o.
  - For GPT-4o, CoT improved inclusion recall but reduced exclusion recall compared with basic prompting.
  - Claude 3.5 Sonnet with CoT achieved balanced performance (N = 530 due to publication processing errors).
- Trade-off between inclusion and exclusion recall was more pronounced with GPT-4o than with Claude 3.5 Sonnet.

Automated prompt optimization

MIPROv2

- AI-modified instructions focused on adapting the general approach rather than specific inclusion/exclusion criteria.
- Compared with basic prompting of GPT-4o, automated prompt optimization improved inclusion recall at the cost of reduced exclusion recall.

Anthropic prompt improver

- Achieved highest exclusion recall across all strategies and maintained inclusion recall performance.
- Small but consistent improvements were obtained over an already high baseline.

Few-shot prompting

- More examples typically improved performance by helping the model better understand the inclusion/exclusion criteria.
- 2-shot prompting was highly specific with high exclusion recall but the trade-off was reduced inclusion recall.
- 4-shot prompting showed optimal performance, with additional examples leading to worse results and significantly increased processing time.

Semantic (K-nearest neighbour [KNN]) few-shot prompting

- 2-shot KNN achieved similar performance to KNN 4-shot prompting.
- Semantic few-shot prompting required fewer examples than few-shot prompting (fewer examples were needed).

MedPrompt framework

- Components tested: GPT-4o with in-context learning + self-generated CoT + temperature = 1.
  - Choice-shuffling ensemble, an additional component of the framework, was not applicable to the binary classification task.
- MedPrompt did not improve performance recall compared with zero-shot prompting or KNN 4-shot prompting.
- 5.5 times longer processing time than zero-shot prompting baseline (165 min vs 30 min).

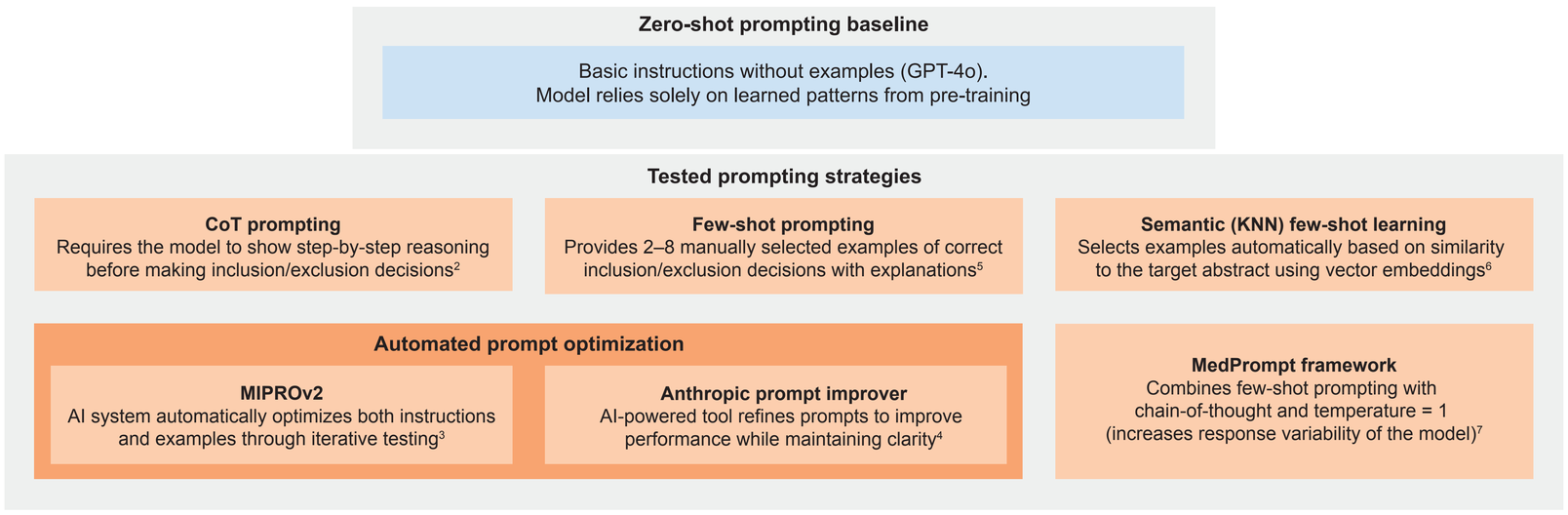

Figure 1. Prompting techniques tested. AI, artificial intelligence; CoT, chain-of-thought; KNN, K-nearest neighbour.

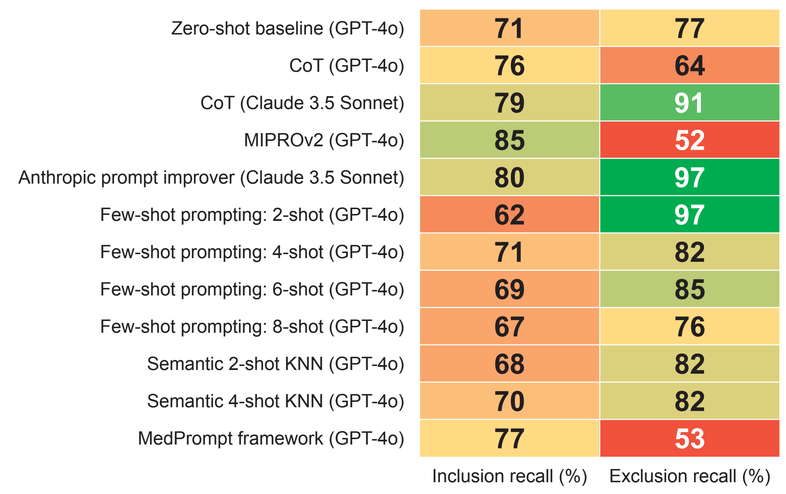

Figure 2. Inclusion and exclusion recall across all models and strategies tested. CoT, chain-of-thought; KNN, K-nearest neighbour.

DISCUSSION

Performance

Inclusion vs exclusion trade-off

- There was generally improved exclusion recall (particularly 2-shot) with few-shot prompting, but at the cost of inclusion recall, suggesting that this approach makes models more conservative in their inclusion decisions.
- For GPT-4o, CoT improved inclusion recall for GPT-4o, but reduced exclusion recall, indicating that this approach may encourage more liberal inclusion decisions when the model explains its own reasoning process.
- Automated optimization tools showed divergent patterns; MIPROv2 favoured inclusion recall improvements while Anthropic prompt improver enhanced both metrics simultaneously, which is critical for maintaining screening quality.
- Semantic 2-shot KNN learning achieved better inclusion recall than traditional 2-shot (62% inclusion) while maintaining high exclusion performance, suggesting that intelligent example selection can reduce the inclusion-exclusion trade-off seen in traditional few-shot approaches.

Model sensitivity

- Findings are consistent with the literature on LLM prompt sensitivity.[8] Small instruction changes significantly impacted performance, reinforcing the need for systematic optimization.

Efficiency considerations

- Semantic few-shot learning was more efficient than manual few-shot learning.
- Medprompt's 5.5-times-longer processing time than zero-shot prompting may limit practical application.
- Semantic 4-shot prompting offers optimal performance–efficiency balance.

Practical implementation

Evidence-based optimization

- This systematic evaluation of multiple prompting strategies provides the methodological rigor needed to deploy AI assistance responsibly, ensuring that technology enhances rather than compromises review quality through data-driven strategy selection.

Optimal strategy selection for reviewers

- High exclusion priority: Anthropic prompt improver with Claude 3.5 Sonnet
- Balanced performance: Claude 3.5 Sonnet with CoT
- Time-constrained: Semantic 4-shot or 2-shot KNN prompting with GPT-4o (good performance, minimal time increase)

Limitations

- This evaluation was conducted using a single systematic review dataset, and focused on a binary include/exclude classification task, with inherently non-deterministic model behaviour.
- LLM screening performance is context-dependent; varying review complexity and scope necessitate tailored prompting strategies.

Conclusions

- The substantial performance variations observed demonstrate that effective AI implementation requires specialized expertise in both systematic review methodology and advanced prompt-engineering techniques.
- Based on this research, we recommend the following strategies for systematic literature review screening:
  - deploy Anthropic prompt improver with Claude 3.5 Sonnet for balanced performance
  - use semantic 4-shot or 2-shot KNN prompting with GPT-4o for time-sensitive projects
  - use a human as a second screener.

References

1. Mikolajewicz N *et al*. *J Bone Miner Res* 2020;35:446–59.
2. LangChain, Inc. Providers: Integration Packages. 2025. Available from: https://python.langchain.com/docs/integrations/providers/ (Accessed 1 October 2025).
3. Opsahl-Ong K *et al*. arXiv:2406.11695 [cs.CL].
4. Anthropic. Improve your prompts in the developer console. 2024. Available from: https://www.anthropic.com/news/prompt-improver (Accessed 1 October 2025).
5. Brown T *et al*. In: Advances in neural information processing systems 33 (NeurIPS 2020). 2024. Available from: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. (Accessed 1 October 2025).
6. Nie F *et al*. arXiv:2212.02216 [cs.CL].
7. Nori H *et al*. arXiv:2311.16452 [cs.CL].
8. Razavi A *et al*. Benchmarking prompt sensitivity in large language models. Proceedings of the European Conference on Information Retrieval, 6–10 April 2025, Lucca, Italy.

Presented at International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Europe 2025, 9–12 November 2025, Glasgow, UK