# How Accurate are Large Language Models for Abstract Screening in Systematic Literature Reviews?

Alparslan Sertel[1], Sumeyye Samur[1], Turgay Ayer[1,2,3], Jagpreet Chhatwal[1,4]

[1] Value Analytics Labs, Boston, Massachusetts, USA; [2] Georgia Institute of Technology, Atlanta, Georgia, USA; [3] Emory Medical School, Atlanta, Georgia, USA; [4] Massachusetts General Hospital Institute for Technology Assessment, Harvard Medical School, Boston, MA, USA.

MSR44

## BACKGROUND

- In healthcare, systematic literature reviews (SLR) are essential for examining specific research topics.
- During screening, human reviewers often assess thousands of articles, which is a time-consuming task prone to errors.
- Recent advancements, like using large language models (LLMs) such as ChatGPT, show potential for automating SLR, especially in text-heavy tasks like screening.
- Prompt engineering can substantially affect the performance of LLMs for conducting SLRs.

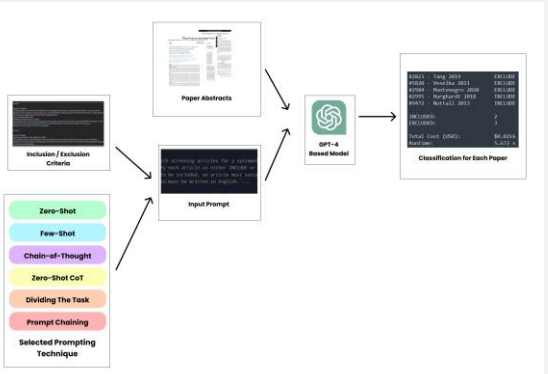
## OBJECTIVE

Our goal was to evaluate the feasibility of conducting abstract screening using LLMs and explore the accuracy of various prompting techniques in the abstract screening step of SLRs. During screening, human reviewers often assess thousands of articles, which is a time-consuming task prone to errors.

## METHODS

- We implemented a binary classification model in Python with the OpenAI GPT-4 API to screen abstracts and categorize them as included or excluded (Figure 1)

**Figure 1:** *Abstract Screening Model Pipeline*




## KEY FINDINGS

- This study sheds light on the effectiveness of five well-known prompting techniques in a conventional abstract screening step of an SLR
- Few-Shot prompting shows the highest performance in terms of accuracy
- The integration of LLMs holds promise in revolutionizing the landscape of SLR However, further improvements are needed to increase the accuracy of LLMs before they can be fully implemented for automating SLRs

- We tested five prompting techniques: Zero-shot, Few-shot, Chain-of-thought (COT), Zero-shot COT, and Dividing into subtasks (Table 1)

**Figure 1:** *Examples for Used Prompt Techniques*

| Prompting Technique | Example Prompt |
|---|---|
| Zero-Shot | "Classify the text into neutral, negative or positive.<br>Text: I think the vacation is okay.<br>Sentiment:" [1] |
| Few-Shot | "A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:<br>We were traveling in Africa and we saw these very cute whatpus.<br>To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:" [2] |
| Chain-of-Thought (COT) | "Question: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can haVE 3 tennis balls. How many tennis balls does he have now?<br>Reasoning: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11.<br>Answer: 11<br>Question: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?" [3] |
| Zero-Shot COT | "Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?<br>A: Let's think step by step." [4] |
| Subtasks | "Please follow these steps:<br>1. Write three topic sentences arguing for {{STATEMENT}}.<br>2. Write three topic sentences arguing against {{STATEMENT}}.<br>3. Write an essay by expanding each topic sentence from Steps 1 and 2, and adding a conclusion to synthesize the arguments. Please enclose the essay in <essay></essay> tags." [5] |

- We employed the LangChain library to conduct parallel testing of these techniques.
- Our experiments included 2,950 papers from initial queries on PubMed and Embase databases.
- Results were compared based on accuracy, defined by percentage of abstracts matching human reviewers' decisions, execution time, and cost.

## RESULTS

- The overall accuracy, encompassing the inclusion and exclusion of abstracts, was highest with Few-shot (82%) and lowest with Chain-of-thought (65%) (Figures 2 and 3).

**Figure 2:** *Overall Accuracy by each Prompting Technique*

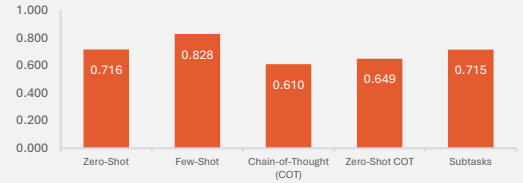
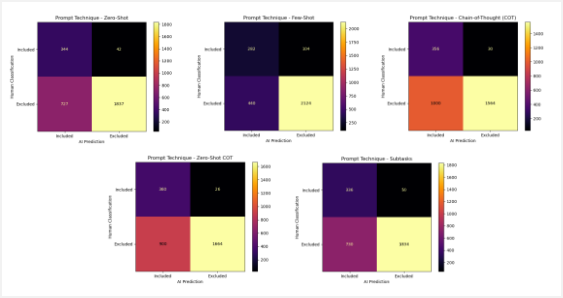
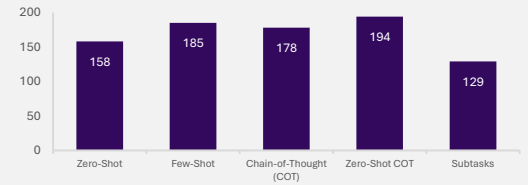



**Figure 3:** *Accuracy for included and excluded papers*



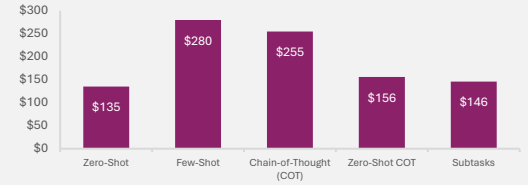
- Although various factors such as the number of tokens, server load, network speed can influence the execution time, in our experiments, Zero-shot was the fastest method whereas the Zero-shot COT was the slowest one (Figure 4).

**Figure 4:** *Execution Time (min) by each Prompting Technique*




- Because cost is primarily linked to the number of tokens in both input and output prompts, the costliest technique was Few-shot ($280) with Zero-shot being the least costly one ($160) (Figure 5).

**Figure 5:** *Cost ($) by each Prompting Technique*




## LIMITATIONS

- Performance of the prompting techniques are evaluated by comparing them with human reviewers' decisions, which are presumed as the gold standard.
- The binary classification model is built over GPT-4 without fine-tuning or additional training, only alteration being the prompting techniques.
- We limited our study to abstract screening step of SLR, and future research should evaluate the performance of full-text review using LLMs.

## REFERENCES

[1] https://www.promptingguide.ai/techniques/zeroshot
[2] https://www.promptingguide.ai/techniques/fewshot
[3] Wei et al. (2022)
[4] Kojima et al. (2022)
[5] https://docs.anthropic.com/claude/docs/break-tasks-into-subtasks

**Corresponding Author:**
Jagpreet Chhatwal, PhD;
JagChhatwal@mgh.harvard.edu