November 2024

White Paper

Bridge
Evidence that matters

Artificial Intelligence in Systematic Literature Reviews
Part 3 | Prompt Engineering

In this white paper we explore the fundamentals of prompt development, extending our previous work on utilizing artificial intelligence (AI) and large language models (LLMs) for title/abstract and full-text screening in systematic literature reviews (SLRs). We define prompts and their basic structures, and then discuss various prompt engineering techniques, from Zero-Shot Prompting (ZSP) to more advanced approaches such as Self-Consistency Prompting (SCP). By providing examples and highlighting the key advantages of each method, we demonstrate how these approaches can enhance the precision and accuracy of AI outputs in SLRs. We also discuss the role of domain expert knowledge in refining prompts, and the importance of selecting appropriate techniques for different stages of the SLR process, including the value of developing bespoke prompts for each SLR.

Introduction

In previous white papers,[1,2] we discussed the use of artificial intelligence (AI) for screening title/abstracts and full-text papers in the development of a systematic literature review (SLR). The AI models discussed were Large Language Models (LLMs), such as those from OpenAI's Generative Pre-Trained Transformer (GPT) series. One of the keys to unlocking the potential of these models lies in the information entered by the user, which is known as the **prompt**.

When we present our detailed research data on AI, or when conducting an AI-enabled SLR with clients, we are frequently asked about prompts. Here we present the basics of prompt development in the context of SLRs; we define prompts, explain the basic prompt structure, and then discuss more sophisticated approaches to prompt engineering.

The reader should keep in mind that this is a fast-moving field, in which novel techniques and enabling tools are evolving rapidly. Therefore, what might be considered 'state of the art' today, will likely be different in a matter of months.

What is a prompt?

In essence, a prompt is a question, a statement, or a more complex set of instructions that provokes a response from an AI model; the response being based on the information and context provided within that prompt. In general, the more specific or refined the prompt, the more useful are the responses from the model. Therefore, **prompt engineering** is the process of designing and refining prompts to get the most accurate and useful responses.

For a straightforward task, a simple prompt will suffice (for example, *"Is the following abstract from a literature review? Answer "yes" or "no"".* However, as task complexity or the amount of data increases, more advanced approaches are required, and the design of such prompts can greatly affect the LLM's output.

In the following section we will discuss different approaches to prompt engineering in the context of title/abstract screening using an example title/abstract (see box). We will explain each approach, demonstrate its implementation, and highlight its key advantages. The following prompting approaches will be covered:

1. Zero-Shot Prompting (ZSP; baseline technique)
2. Few-Shot Prompting (FSP)
3. Chain of Thought Prompting (CoTP)
4. Clues and Advanced Reasoning Prompting (CARP)
5. Self-Consistency Prompting (SCP)

Prompting Techniques

We use ZSP as the **baseline technique** against which we compare the more advanced techniques. With ZSP, the prompt provides minimal context for the LLM, which must generate an appropriate response based solely on its prior knowledge.[3]

No additional task-specific data or instruction is provided. **Table 1** provides an example of a ZSP prompt used to classify a title/abstract as primary study or review, and the response from the LLM.–

Both **Table 1** and **Table 3** show prompts and corresponding responses generated by the GPT-4 model. In order to keep the tables relatively brief, we have truncated some of the prompts and responses. The hyperlinks in the both table allow the reader to view each prompt and response in the ChatGPT interface (**please note** that if you paste in the example title/abstract and prompts, rather than clicking on the hyperlinks, you may see a slightly different response; LLM outputs can vary based on several other parameters such as the model used, model version, model temperature etc.).

Example title/abstract for demonstrating various prompting techniques

Title: Quality of Life in Patients with Advanced Non-Small Cell Lung Cancer: A WHOQOL-Based Analysis Using US Healthcare Data

Background: NSCLC has an incidence of 45.4 per 100,000. It significantly impairs patient quality of life (QoL) due to its advanced stage at diagnosis and aggressive disease progression. This study aims to assess the QoL of patients with advanced stages of non-small cell lung cancer (NSCLC) using the World Health Organization Quality of Life (WHOQOL) scale through a comprehensive analysis of US healthcare data.

Methods: Patients aged 18 years and older, diagnosed with NSCLC at advanced stages between 2019 and 2020 and treated with chemotherapy, were included in the study. QoL was assessed using the WHOQOL-BREF questionnaire, covering physical health, psychological health, social relationships, and environmental domains. Descriptive statistics summarized patient characteristics and QoL scores. Multivariate linear regression models were utilized to identify factors associated with QoL outcomes.

Results: A total of 8,892 patients with late-stage NSCLC were enrolled, of which 6,536 had advanced (Stage III) disease. Common chemotherapy treatments used in these patients included cisplatin, carboplatin, paclitaxel, docetaxel, and pemetrexed. In patients with Stage IV NSCLC, the mean overall WHOQOL-BREF score was 52.3 (standard deviation [SD] ± 12.7), indicating moderate QoL impairment. Physical health and psychological health domains had the lowest scores, with means of 48.1 and 50.2, respectively. Factors significantly associated with lower QoL included advanced age ($\beta = -0.45$, $p<0.001$), presence of comorbidities ($\beta = -0.62$, $p<0.001$), and lower socioeconomic status ($\beta = -0.38$, $p=0.002$). Over 1 year, the mean WHOQOL-BREF score improved by 2.3 points (SD ± 1.7).

Conclusion: Patients with advanced NSCLC treated with chemotherapy experience considerable QoL impairments, particularly in the physical and psychological domains. Further research is needed to understand impaired QoL in the context of disease progression.

[Note: This is a fictional title/abstract. In real-world use, the prompts we develop for each client project are used across thousands of title/abstracts. In order to demonstrate the benefits of advanced prompting approaches in handling complexity and ambiguity, we have created this fictional title and abstract to represent information that might be found across multiple title/abstracts]

Table 1: Zero-Shot Prompting for classifying a title/abstract

Technique	Zero-Shot Prompting (ZSP)
ZSP Prompt	Based on the title/abstract, is this a primary study or a literature review?
LLM response to ZSP Prompt	Based on the information provided, this appears to be a **primary study**. The document contains specific sections typically found in a primary research article, such as Background, Methods, Results, and Conclusion. It mentions a particular study that has been carried out, including details about the study's methodology (use of WHOQOL-BREF questionnaire), the population studied (patients with advanced NSCLC), and statistical analyses (descriptive statistics and multivariate linear regression). The presence of these details, particularly the original data and analyses reported in the Results section, indicates that this is original research rather than a review of literature. [Link to LLM response]

ZSP can adequately address such simple queries; however, in the case of more complex or less common queries, ZSP may result in superficial, vague, and ambiguous answers.

Table 2 summarizes the more advanced prompting techniques used with LLMs, briefly explaining their key features. Each of these techniques offers unique benefits: FSP enhances precision and accuracy by providing contextual examples; CoTP improves the model's ability to handle complex tasks through step-by-step reasoning; CARP focuses the model on critical information for detailed analysis; and SCP increases reliability and factual consistency by comparing multiple reasoning pathways.[4-6]

Table 2: Key features of more advanced prompting techniques for more complex data queries

Technique	Brief description
Few-Shot Prompting (FSP)	FSP enhances an AI language model's performance by including a small number of task-specific examples within the prompt, providing concrete guidance. This method improves the model's understanding, **reduces ambiguity**, and leads to more accurate and **consistent outputs** compared with ZSP, which relies solely on general instructions without examples. By learning from these specific instances, the model aligns more closely with the desired output format, making FSP particularly advantageous over ZSP in terms of precision and reliability.
Chain of Thought Prompting (CoTP)	CoTP enhances an AI language model's reasoning capabilities by encouraging it to generate intermediate reasoning steps that lead to the final answer. By incorporating explicit examples of thought processes or instructing the model to "think step-by-step" within the prompt, CoTP enables the model to tackle **complex problems** more effectively than ZSP, which provides answers without guided reasoning. This improves accuracy and coherence by making the reasoning process **transparent**, resulting in more reliable and interpretable outputs.
Clues and Reasoning Prompting (CARP)	CARP enhances an AI language model's performance by incorporating specific clues and encouraging detailed reasoning within the prompt. This method guides the model through a **structured problem-solving process**, leading to more **accurate and coherent responses** compared with ZSP, which lacks such guidance. By providing clues and fostering step-by-step reasoning, CARP reduces ambiguity and improves the model's ability to handle complex tasks, resulting in outputs that are both **precise and reliable**.
Self-consistency Prompting (SCP)	SCP improves an AI language model's accuracy by generating multiple reasoning paths for a given problem and selecting the most consistent answer among them. This method leverages the model's ability to explore different solutions, **reducing errors** that might occur in a single attempt, as seen in ZSP which provides only one immediate response without self-evaluation. By comparing and aggregating these multiple outputs, SCP enhances **reliability and precision**, leading to more accurate and robust results.

Table 3 illustrates how these approaches can lead to more precise and accurate responses compared with basic ZSP, with reference to the example title/abstract shown earlier. As mentioned above, both the prompts and responses have been truncated for the table, but full information can be found via the links in each relevant cell.

Table 3: Examples of differences between basic ZSP prompts and more advanced prompts

Parameter	ZSP prompt/ Research question	LLM response to ZSP prompt	Advanced prompt	LLM response to advanced prompt (FSP, CoTP, CARP, SCP)	Advantage of advanced prompt over ZSP prompt
Study design	**ZSP**		**FSP**		• **Increased precision**: FSP prompting led the LLM to identify the study as a *retrospective cohort study*, while ZSP prompting resulted in a more general classification as an *observational cohort*. • **Improved format compliance**: With an FSP prompt, the LLM responded in exactly the desired format, stating only the study design without extra text or explanation.
	Identify the study design of this title/abstract.	The response explains that the study is an **observational cohort study** based on its patient sample, non-interventional analysis etc. [Link to LLM response]	Identify the study design and provide the answer for study design in the exact format given in the examples.	Study design: **Retrospective cohort study**. [Link to LLM response]	
Disease	**ZSP**		**CoTP**		• **Increased precision**: CoTP prompting led the LLM to specifically identify the disease of interest as *advanced* NSCLC, whereas ZSP prompting resulted in a more general identification as NSCLC.
	Identify the disease of interest in this title/abstract.	The disease of interest in the provided title and abstract is **non-small cell lung cancer (NSCLC).** [Link to LLM response]	Think step-by-step to identify the disease of interest from the title/abstract. Steps: Disease of interest:	The response outlines the five steps taken by the LLM to identify the disease of interest, concluding that it is **advanced non-small cell lung cancer**. [Link to LLM response]	
Outcomes	**ZSP**		**CARP**		• **Increased specificity**: CARP prompting led the LLM to correctly identify QoL as the *only* outcome of interest, whereas ZSP prompting incorrectly mentioned that the abstract also provides insight into treatment patterns.
	Which of the following outcomes are reported in the abstract: incidence, prevalence, QoL, and others?	The response lists the study's reported outcomes, highlighting **QoL** as the primary focus, mentions **treatment patterns**, and notes that other outcomes are not explicitly reported. [Link to LLM response]	The prompt instructs the LLM to identify relevant outcomes by finding clues, reasoning analytically, and determining the reported outcomes.	The response lists clues from the title and abstract, uses reasoning to analyze them, and decides that **QoL** is the primary outcome reported. [Link to LLM response]	
Sample size	**ZSP**		**SCP**		• **Improved accuracy**: SCP prompting led the LLM to provide an accurate sample size, whereas ZSP prompting resulted in an incorrect estimation.
	What is the metastatic NSCLC sample size in this title/abstract?	The sample size for patients with mNSCLC in this study is **8,892.** [Link to LLM response]	Using self-consistency, determine the sample size of mNSCLC. Present three different interpretations of the data and then identify the most consistent sample size.	The response analyzes the data to deduce that **2,356** patients are mNSCLC patients, interpreting the total minus Stage III patients as Stage IV (metastatic), and concludes this is the most consistent interpretation. [Link to LLM response]	

CARP: Clues and Reasoning Prompting; CoTP: Chain of Thought Prompting; FSP: Few-Shot Prompting; LLM: Large Language Model; mNSCLC: metastatic Non-Small Cell Lung Cancer; NSCLC: Non-Small Cell Lung Cancer; QoL: Quality of Life; SCP: Self-Consistency Prompting; ZSP: Zero-Shot Prompting.

Discussion

In this paper, we have provided a brief introduction to some prompt engineering techniques that may be used in working with LLMs. While the examples we discuss demonstrate each method used independently and in isolation, in practice, these techniques can be combined into 'hybrid prompts' for greater effectiveness. Thus, CoTP may be combined with FSP to ensure that accurate outputs are obtained (through CoTP), and in the desired format (through FSP). For tasks requiring high accuracy and depth of understanding, 'prompt chaining' may be used, in which data are passed between multiple LLMs, leveraging their individual strengths and reducing the impact of any single model's limitations. Use of 'agents' (software that acts as intermediaries to manage prompts, interpret responses, and optimize advanced techniques) can further improve LLM outputs by making them more accurate and context-aware. We will elaborate on these more recent techniques in a future paper.

From our experience testing various LLMs across the different stages of producing an SLR, we have observed that using more complex prompt techniques does not always lead to higher accuracy. The key to improving accuracy is selecting the *appropriate* prompt technique for the specific task at hand, whether title/abstract screening, full-text screening, data extraction, or data summarization. It is also clear that expert knowledge is vital for refining and improving prompts.

For example, in burden-of-illness reviews involving observational study data, we found that LLMs often fail to differentiate between studies whose primary focus is the disease of interest and others which report it only as a comorbidity of another condition. During title/abstract screening, such errors can lead to a large number of false positives. In these situations, subject matter experts can identify flaws in the reasoning steps followed by LLMs and correct them with improved prompts. Experts can also provide key domain-specific clues to guide the LLMs. Therefore, even when using the same prompt *technique*, the accuracy of outputs can improve substantially with expertly tailored prompts that are specific to individual reviews and research questions.

For these reasons, we construct bespoke (i.e., customized) prompts for each stage of each AI-enabled SLR we conduct. The additional work requirements for fine-tuning our library of prompts have been offset by developing a software stack that – among other things – allows for rapid testing of multiple prompt approaches until we see a level of performance equivalent to that of a highly experienced outcomes researcher. Furthermore, as models evolve, so too will the need for continual prompt refinement. One benefit of the latest GPT model at the time of writing [o1] is that this model is trained to use CoTP,[7] which allows the user to follow the logical 'thinking' process of the LLM.

While prompt engineering is vital to working with LLMs, other factors also play a role. One can improve LLM performance by giving the AI a specific role or persona, providing clear context, and using tools such as text retrieval systems. In addition, modern Application Programming Interfaces (APIs) for LLMs allow developers to automate tasks using code instead of performing them manually. For example, they can sequentially feed in a large number of title/abstracts in an automated manner. These APIs also provide adjustable settings to fine-tune the model's behavior for specific goals. For example, developers can modify the model's temperature.[8]

In summary, prompt techniques are central to maximizing the benefits of LLMs in SLRs. Selecting the correct technique for the context is important, and domain expertise remains key to prompt refinement. Our experience suggests that bespoke prompts are required to achieve maximum benefit for each SLR, and for each stage of each SLR. To this end, automation of prompt testing and refinement at scale will become an increasingly important component in the implementation of high-quality AI-enabled SLRs.

Authors: Saifuddin Kharawala, Sam Isaacs, Paul Gandhi

References

1. Artificial Intelligence in Systematic Literature Reviews Part 1 | AI-aided Title/Abstract Screening. https://www.bridgemedical.org/site/assets/files/2156/ai_in_literature_reviews_white_paper_31_jan_24-1.pdf. Accessed 04 November 2024.
2. Artificial Intelligence in Systematic Literature Reviews Part 2 | AI-aided Full-text Screening. https://www.bridgemedical.org/site/assets/files/2202/white_paper_on_ai-aided_full-text_screening_in_slrs_23_august_2024-1.pdf. Accessed 04 November 2024.
3. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., 2022. Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916.
4. Jain, S., Ma, X., Deoras, A. and Xiang, B., 2023. Self-consistency for open-ended generations. arXiv preprint arXiv:2307.06857.
5. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T. and Wang, G., 2023. Text classification via large language models. arXiv preprint arXiv:2305.08377.
6. Wang, X., Jason W., Dale S., Quoc L., Ed C., Sharan N., Aakanksha C. and Denny Z., 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.
7. https://medium.com/@saman.rahbar/a-comprehensive-technical-review-of-openais-gpt-o1-and-transformer-advancements-021945a10b0d. Accessed 04 November 2024
8. https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results. Accessed 04 November 2024.