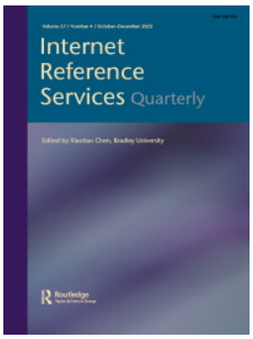

Internet Reference Services Quarterly

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/wirs20

The Art and Science of Prompt Engineering: A New Literacy in the Information Age

Leo S. Lo

To cite this article: Leo S. Lo (2023) The Art and Science of Prompt Engineering: A New Literacy in the Information Age, Internet Reference Services Quarterly, 27:4, 203-210, DOI: 10.1080/10875301.2023.2227621

To link to this article: https://doi.org/10.1080/10875301.2023.2227621

INTERNET REFERENCE SERVICES QUARTERLY
2023, VOL. 27, NO. 4, 203–210
https://doi.org/10.1080/10875301.2023.2227621

Routledge
Taylor & Francis Group

COLUMN

The Art and Science of Prompt Engineering: A New Literacy in the Information Age

Leo S. Lo

University of New Mexico, Albuquerque, New Mexico, USA

ABSTRACT

The novel discipline of prompt-engineering is a combination of artificial intelligence, linguistics, and user experience design. Crafting effective prompts for AI models like OpenAI's GPT can optimize the quality of generated output, especially in reference services. The article explains the CLEAR Framework—a guiding tool that incorporates principles of conciseness, logic, explicitness, adaptability, and reflectiveness in prompt crafting. Despite challenges including biases, ethical concerns, and keeping up with rapidly evolving AI capabilities, the field also presents opportunities for growth. The article concludes with a call to action for library professionals to embrace and master prompt-engineering as an essential skill.

KEYWORDS

Prompt engineering; artificial intelligence; ChatGPT; AI-assisted services; generative AI

Introduction

In the era of pervasive artificial intelligence (AI), we are witnessing transformative changes across numerous industries and fields. The library and information services sector is no exception. Advanced AI models like OpenAI's GPT have started to play a more substantial role, creating opportunities for optimized services and resources (Boyd, 2023; Khan, 2023). However, effectively harnessing these AI models necessitates understanding prompt engineering—the art and science of designing effective input instructions. This column will delve into the world of prompt engineering, outlining the technical principles and creative strategies that form its foundation, while exploring the role it can play in library and information services.

Understanding GPT: The journey from prompt to response and the science of prompt engineering

GPT models are founded on the Transformer architecture, which utilizes self-attention mechanisms and multi-layer feed-forward neural networks. These components work in concert to process and generate text, allowing GPT models

CONTACT Leo S. Lo leolo@unm.edu University of New Mexico, Albuquerque, NM 87131, USA.

to respond in a way that is contextually meaningful. Trained on vast volumes of text data, GPT learns to capture complex linguistic patterns and relationships, making it a powerful tool for a variety of NLP tasks, including machine translation, summarization, and conversational AI (Brown et al., 2020).

The journey of a prompt within a GPT model can be broken down into a series of steps:

1. Tokenization: The model starts by converting the input prompt into a sequence of tokens, often in the form of subword units, using a process known as Byte-Pair Encoding (BPE).
2. Context Window Selection: Should the prompt exceed GPT's maximum token limit (for example, GPT-3 has a 4096-token limit), it must be truncated or divided into smaller parts, risking a potential loss of context.
3. Embedding: Each token is then mapped to a vector representation, called an embedding. This captures semantic and syntactic information about the token.
4. Forward Pass: The embeddings are processed through a series of Transformer blocks. These leverage self-attention mechanisms and neural networks to identify long-range dependencies and contextual relationships between tokens.
5. Probability Distribution: GPT then produces a probability distribution for the next token in the sequence. This predicts possible next words based on the patterns it learned during training.
6. Token Sampling: A token is chosen from the probability distribution. The method of selection, such as greedy or temperature-based sampling, can influence the randomness and diversity of the text generated.
7. Decoding: The chosen token is added to the existing sequence, and the process is repeated until a specific criterion for stopping is met, which might be a maximum token limit, an end-of-sentence token, or a condition defined by the user.
8. Detokenization: Finally, the sequence of tokens is translated back into human-readable text, with any special tokens used during tokenization being removed and original words or subwords being reconstructed.

The efficiency of this journey from prompt to response is significantly influenced by the science of prompt engineering. This involves the construction of prompts that effectively guide the AI model's processing and generation of text.

Prompt engineering revolves around several key principles:

1. Clarity and Precision: Clear and precise prompts guide the AI model's understanding, impacting tokenization and embedding. They help ensure accurate embeddings are generated, leading to more relevant responses.

2. Contextual Information: Providing context within prompts helps the AI model generate accurate and contextually fitting responses. Contextual information assists the model in forming meaningful connections between tokens, improving its overall understanding and the quality of the generated output.
3. Desired Format: Prompt engineering can indicate the expected response format. This guidance assists the AI model in generating output that meets user expectations and follows a specific structure or style.
4. Verbosity Control: Managing the length of the generated response is another aspect of prompt engineering. Instructions within the prompt can encourage the AI to produce concise or detailed responses, allowing for greater control over the output's verbosity.

These principles form the foundation of effective prompt engineering. They directly influence the performance of generative AI models. By understanding the technical foundations of these principles and applying them in practice, we can optimize the interaction between humans and AI. This results in more accurate, relevant, and contextually appropriate responses across a broad spectrum of NLP tasks, underlining the powerful capability of GPT models.

The artistry in prompt engineering

While the technical principles form the bedrock of prompt engineering, the nuances of creativity, intuition, and iterative refinement infuse it with a distinct artistic flair. It is this blend of science and art that empowers prompt engineers to craft more engaging, contextually rich, and effective prompts.

Creativity

Creativity is a cornerstone in the art of crafting effective prompts. This involves innovative thinking and strategic experimentation with different styles, tones, and structures of prompts. In the context of a library AI chatbot, patrons may hail from diverse backgrounds, boasting different levels of proficiency and comfort with language. They may use varying styles, ranging from formal to colloquial, to interact with the chatbot. Hence, experimenting with different phrasings, introducing analogies or examples, and even adopting more conversational styles can help to create prompts that yield engaging, satisfactory, and user-aligned AI-generated responses.

Intuition

Intuition plays a key role in prompt engineering. Developing an understanding of how AI models respond to different prompts comes with

experience, observation, and thoughtful analysis. Understanding the subtle nuances and patterns in AI responses, such as the impact of framing the prompt as a question versus a statement, or the strategic use of specific keywords or phrasing, can significantly influence the quality, relevance, and coherence of AI-generated responses.

Iterative refinement

Iterative refinement is an essential aspect of the art of prompt engineering. It involves a continuous process of testing prompts, analyzing the generated AI responses, and making necessary adjustments based on observed patterns and outcomes. This iterative feedback loop allows for constant learning and refinement of the prompts and consequently, the enhancement of AI model performance and the optimization of the quality and relevance of AI-generated content.

Personalization and adaptation

Prompt engineering also necessitates personalizing prompts to cater to individual user preferences and needs. Personalization might involve modifying the prompt's tone, style, or level of detail to resonate with individual users. For instance, prompts can be tailored to match the linguistic style of a user or can be designed to incorporate more domain-specific language for users seeking expert-level information. This personal touch in prompt engineering fosters more meaningful, engaging, and satisfying human-AI interactions.

Collaboration and interdisciplinary knowledge

Prompt engineering is not an isolated discipline but a collaborative field that benefits greatly from interdisciplinary knowledge and expertise. For instance, partnering with subject matter experts can help prompt engineers craft more accurate, relevant, and context-rich prompts for specialized fields. Simultaneously, collaboration with linguists or psychologists can offer invaluable insights into the cultural, cognitive, and social intricacies of language. These insights can further enrich prompt design, enabling the creation of prompts that are sensitive to cultural nuances, cognizant of user cognition, and mindful of social context. This symbiosis of different fields and expertise adds another layer of depth and effectiveness to the art of prompt engineering.

Integrating the CLEAR framework into the art and science of prompt engineering

Building on the intricate interplay of scientific methods and artistic touches in prompt engineering, a unified, structured approach is vital to maximize

the effectiveness of AI language models like ChatGPT in academic libraries. The CLEAR Framework is a guide for this purpose, centering around five critical elements—Concise, Logical, Explicit, Adaptive, and Reflective (Lo, 2023).

Each component of the CLEAR Framework caters to a distinct aspect of prompt engineering, arming users with a comprehensive set of techniques to enhance their engagement with AI language models. The potency of the framework arises from its methodical approach to common hurdles in generative AI, like generating cohesive and context-appropriate responses. Crafted with flexibility at its core, the CLEAR Framework allows users to mold its principles according to different AI language models and educational scenarios.

Elements of the CLEAR framework

Concise: Brevity and clarity in promptsThe principle of conciseness underscores the importance of succinct and clear prompts. Such prompts guide AI language models to concentrate on the core elements of the task, thereby generating more relevant and precise responses.

Example: Instead of, "Could you delve into the stages of mitosis and why it's significant for cell division?," go for a concise version, "Describe the stages of mitosis and their importance in cell division."

Logical: Structured and coherent prompts

Logical prompts focus on preserving a logical progression of ideas within a prompt. A logically crafted prompt enables AI models to grasp the context and the interconnections between various concepts, leading to more coherent and precise outputs.

Example: A logically structured prompt like, "Outline the steps involved in creating a financial budget, beginning with income assessment and concluding with expense tracking," can yield an orderly and complete response.

Explicit: Clear output specifications

The principle of being explicit underlines the need for precise output specifications in the queries. Clearly stating the expected output format, content, or scope minimizes the risk of unsuitable or irrelevant responses from the AI model.

Example: Instead of asking, "Can you tell me about the Civil War?," an explicit prompt would be, "Give a brief summary of the American Civil War, concentrating on its origins, key battles, and aftermath."

Adaptive: Flexibility and customization in prompts

Adaptability emphasizes the need for flexibility in prompt engineering. It entails experimenting with a variety of prompt formations, phrasings, and temperature settings to find a balance between creativity and specificity.

Example: If a general prompt like, "Discuss the effects of global warming," results in broad responses, a more targeted and adaptive prompt might be, "Analyze the impact of global warming on polar ice caps."

Reflective: Continuous evaluation and improvement of prompts

Reflectiveness encourages ongoing assessment and refinement of prompt engineering techniques. Through careful evaluation of AI-generated responses, based on factors such as accuracy, coherence, and utility, users can improve their future prompts and increase the effectiveness of their AI engagements.

Example: After receiving a set of AI-generated responses, conduct an in-depth review and apply these insights for crafting more effective future prompts.

To sum up, the CLEAR Framework offers a solid scaffold for devising effective prompts for AI language models, combining precision and creativity, science and art.

Challenges and opportunities in prompt engineering

Even though prompt engineering has the potential to revolutionize the way we interact with AI, it is not immune to challenges. These hurdles span various domains, from addressing biases and ethical concerns to managing specialized subject matter and adapting to the ever-evolving AI capabilities. Nevertheless, each challenge invariably offers opportunities for innovation, learning, and growth, fueling the continuous evolution of the field.

Addressing biases and ethical concerns

AI models, including language models, have the potential to reflect and amplify biases present in their training data. Bias in AI-generated content can have serious implications, potentially perpetuating harmful stereotypes or providing misleading information. Further, there are ethical concerns to consider, such as the model's potential to generate inappropriate or harmful content, despite efforts to instill safeguards.

Opportunity: The challenge of addressing biases and ethical concerns presents the opportunity to develop fairer, more ethical AI systems. Researchers and engineers can work to improve the methods used to train these models, using techniques like data de-biasing and differential privacy. Also, the integration of ethical considerations into prompt engineering can foster the development of more responsible AI tools, encouraging inclusivity and fairness in AI interactions.

Handling specialized domains

Many applications of AI language models involve specialized domains that require specific knowledge, such as law, medicine, or academia. Crafting effective prompts in these areas can be challenging due to the complexity and the need for precision in these fields.

Opportunity

The difficulty of managing specialized domains presents a unique opportunity to enhance the accuracy and relevance of AI-generated content. Collaborating with subject matter experts can lead to the development of more accurate and contextually relevant prompts. This interdisciplinary approach to prompt engineering can ultimately improve the quality and usefulness of the AI responses in these fields.

Adapting to evolving AI capabilities

AI is a rapidly evolving field, with models continuously growing in size and complexity. Staying abreast of these advancements and understanding how to optimize prompt engineering for new models can be a significant challenge.

Opportunity

On the bright side, the evolution of AI capabilities paves the way for innovation in prompt engineering strategies. As AI models become more advanced, they may be able to understand more complex prompts and produce more accurate and creative outputs. This constant evolution offers the opportunity for continuous learning and growth, pushing the boundaries of what is possible with AI-generated content.

Conclusion and call to action

As we usher in a future increasingly intertwined with generative AI models, mastering the art and science of prompt engineering is not just beneficial—it's essential. Whether it is in understanding the underlying technical principles, implementing creative strategies, iterating on best practices, or exploring the potential of the CLEAR framework, prompt engineering holds the key to optimizing AI-generated content for an array of tasks, particularly within the library and information services sector.

These advancements in AI aren't a passing trend but a revolutionary force reshaping our ways of managing and delivering information services. As such, professionals in the library and information sector must not only stay updated with these developments but also actively immerse themselves in learning and utilizing the skills of prompt engineering. Recognizing and harnessing the potential of prompt engineering can turn AI models from generic tools into highly specialized assets that can cater to a variety of information needs.

This is not merely a call to adaptation but an invitation to revolutionize the way we think about AI and information services. By embracing prompt engineering, we stand on the precipice of a new era where AI is not just a tool but a powerful ally, fine-tuned and optimized to enhance our work

in the most meaningful ways. The journey of mastering prompt engineering may be complex and fraught with challenges, but the rewards are monumental. So, let us seize this opportunity to become pioneers in the evolving landscape of AI-assisted library services. Let us harness the power of AI, guided by our expertise in prompt engineering, to create a future where our library services are not just efficient, but extraordinary.

This is our call to action—let's transform the role of AI in our libraries, optimize our services, and forge a future that exemplifies the pinnacle of what AI and human ingenuity can achieve together.

References

Boyd, E. (2023, May 11). ChatGPT is now available in Azure OpenAI Service. *Azure blog: Microsoft Azure*. https://azure.microsoft.com/en-us/blog/chatgpt-is-now-available-in-azure-openai-service/.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. doi:10.48550/arXiv.2005.14165

Khan, S. (2023, March 15). *Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access!* Khan Academy. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/.

Lo, L. S. (2023). The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. doi:10.1016/j.acalib.2023.102720