IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023

DOI: 10.17148/IJARCCE.2023.121117

Advancements in Prompt Engineering: A Comprehensive Survey

Sumaiya. M. S[1], Yaseen Dada Shaik[2], Sonia Maria Dsouza[3]

Department of Machine Learning B.M.S. College of Engineering, Bangalore, India[1,2,3]

Abstract: This survey paper provides a thorough examination of the rapidly evolving field of prompt engineering, a crucial aspect of natural language processing and artificial intelligence. Prompt engineering involves crafting effective instructions or queries to elicit desired responses from language models. The paper begins by elucidating the foundational concepts of prompt engineering, exploring its historical development, and presenting key methodologies employed in generating prompts for various language models.

By synthesizing existing knowledge and highlighting emerging trends, this paper aims to provide researchers, practitioners, and enthusiasts with a comprehensive understanding of the current state and future directions of prompt engineering in natural language processing.

I. INTRODUCTION

In the realm of natural language processing (NLP) and artificial intelligence (AI), the role of effective communication between humans and machines is paramount. Prompt engineering, a burgeoning field at the intersection of linguistics and machine learning, plays a pivotal role in shaping this interaction. This comprehensive survey paper delves into the nuanced landscape of prompt engineering, a discipline that revolves around crafting precise and contextually rich instructions to extract desired responses from language models. As AI systems have grown in complexity and capability, the importance of formulating prompts that can harness the full potential of these models has become increasingly apparent. The objective of this survey is to provide an extensive exploration of prompt engineering, tracing its historical evolution, elucidating fundamental concepts, and examining the latest methodologies that have emerged to enhance the performance of language models across diverse applications.

The initial sections of this paper lay the groundwork by elucidating the foundational principles of prompt engineering. Subsequently, the survey delves into the methodologies employed in generating effective prompts for language models. Through this exploration, the paper seeks to offer readers a comprehensive understanding of the underlying techniques that contribute to the success of prompt engineering.

As the survey progresses, it transitions to a detailed review of recent advancements in prompt design strategies. The synthesis of this knowledge not only serves as a valuable resource for researchers and practitioners but also contributes to the ongoing discourse on the evolving landscape of prompt engineering in NLP and AI. Ultimately, this survey endeavours to bridge the gap between theory and application, offering a holistic view of prompt engineering's current state while paving the way for future innovations in human-machine communication.

II. LITERATURE SURVEY

1. Prompt engineering guidelines for LLMs in Requirements Engineering

This paper is published in June,2023. Prompt engineering guidelines for how to utilize large generative AI models in the field of requirements engineering are limited in the literature. The utilization of AI for RE as a field has made a lot of progress in the past decades, especially since the introduction of NLP with the use of machine learning and deep learning which in NLP4RE'18 was mentioned to facilitate utilization of NLP tools and techniques. The possible usage of Prompt engineering guidelines within the domain of RE, as well as what advantages and limitations they may introduce are explored.

The 5 databases used for this review were ACM, Scopus, IEEE Explore, Science Direct, and arXiv.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023
DOI: 10.17148/IJARCCE.2023.121117

The guidelines include Context, Template, Persona, Disambiguation, Reasoning, Analysis, Keywords, Wordings, Shorten, Few-shot prompts.

The results are categorized into 3 subsections:

A. The first subsection, subsection A, presents their findings, its' details, and answers for RQ1 with a non-exhaustive list of PE guidelines,

B. In the second subsection, subsection B, they have presented their findings and answer RQ2 by suggesting a mapping of guidelines and themes from RQ1 to various components within RE activities of similar nature.

C. The third subsection, subsection C, presents possible advantages and limitations when applying the PE guidelines in RE.

2. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering

This paper was published on 27th June, 2023. There has been a growing effort to replace hand extraction of data from research papers with automated data extraction based on natural language processing, language models, and recently, large language models (LLMs). In this work the ChatExtract method is proposed that can fully automate very accurate data extraction with minimal initial effort and background, using an advanced conversational LLM.

In this paper it is demonstrated that using conversational LLMs such as ChatGPT in a zeroshot fashion with a well-engineered set of prompts can be a flexible, accurate and efficient method of extraction of materials properties in the form of the triplet Material, Value, Unit. The data extraction is done in two main stages:

1. Initial classification with a simple relevancy prompt, which is applied to all sentences to weed out those that do not contain data.
2. A series of prompts that control the data extraction from the sentences categorized in stage (A) as positive (i.e., as relevant to the materials data at hand). The working of stage B is done as:

i) Split data into single- and multi-valued because single-valued doesn't require much follow-up prompts and multi-valued more prone to errors and requires further scrutinizing and verification. ii) Include explicitly the possibility that a piece of the data may be missing from the text to avoid model from hallucinating. iii) Enforce a strict Yes/No format of answers to reduce uncertainty and allow for easier automation. The extracted databases of critical cooling rates and yield strength for HEAs, data used in the assessment of the models is available on figshare.

A series of prompts were given and a result of 90% precision at 87.7% recall on their test set of bulk modulus data, and 91.6% precision and 83.6% on a full database of critical cooling rates were obtained. They have further developed two databases using ChatExtract - a database of critical cooling rates for metallic glasses and yield strengths for high entropy alloys.

3. Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models

This paper was published on 16th of August, 2022.Neural language models can now be used to solve ad-hoc language tasks through zero-shot prompting without the need for supervised training. PromptIDE allows users to experiment with prompt variations, visualize prompt performance, and iteratively optimize prompts. They have developed a workflow that allows users to first focus on model feedback using small data before moving on to a large data regime. The tool then allows easy deployment of the newly created ad-hoc models.

The main elements of prompting can be summarized as:

1) M1 - Prompt Template. A user writes a prompt template consisting of a task description in natural language that utilizes the fields from the task in a situated context. This leads to the construction of the input x that is used for conditioning of the LLM. 2) M2 - Answer Choices. A user provides a dictionary of answer choices paired with the original labels for a given task that offer different possible output wordings y to be considered by the model. The underlying model uses ranking to determine which of these answer choices to select. The original label paired with this answer choice is then the classification choice selected.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023
DOI: 10.17148/IJARCCE.2023.121117

3) M3 - Evaluation. A user can evaluate the current version of the system under a known prompt for a set of validation data. This step will provide a proxy score for how well the given wording of the prompt is at capturing the underlying task.

Use cases:

1) Documentation Classification
2) Multiple-choice answering
3) Sentence similarity

They have presented PromptIDE, a system for domain experts to customize models for ad-hoc tasks without requiring training expertise.

4. Solving Probability and Statistics Problems by Program Synthesis

This paper was published on 16th November, 2021. They have solved university level probability and statistics questions by program synthesis using OpenAI's Codex, a Transformer trained on text and fine-tuned on code. Their approach requires prompt engineering to transform the question from its original form to an explicit, tractable form that results in a correct program and solution. This work is the first to introduce a new dataset of universitylevel probability and statistics problems and solve these problems in a scalable fashion using the program synthesis capabilities of large language models. The key to success lies in the carefully engineered prompts we present to Codex.

- Foundation models :

For the task of answering questions specifically, such models have recently achieved strong performance. However, when tasked with solving university-level quantitative problems, foundation model performance is poor.

- Probability benchmarks

Datasets, such as MATH, MAWPS, MathQA, Math23k, and GSM8K focusses on benchmarking mathematical question answering, including probability questions, but all of these works only consider grade-school level question difficulty.

1) The first dataset consists of applied questions in probability. 20 questions are taken that have numerical answers from MIT's 18.05: Introduction to Probability and Statistics.
2) The second dataset, 20 questions are taken that have numerical answers from Harvard's STAT110: Probability and

Brainstellar.

Three classes of prompts that are communicated to Codex with and their associated effects are:

1) Program task specification
2) Probabilistic simulation programming
3) Concept grounded task

There is an average similarity of 0.80 in MIT's 18.05 and an average similarity of 0.79 in STAT110.

III. METHODOLOGY

I. Research questions and objectives The possible usage of Prompt engineering guidelines within the domain of Requirements engineering, as well as what advantages and limitations they may introduce are explored. In order to realize these objectives, the following research questions were formed:

RQ1: What does the existing literature regarding PE guidelines for large generative AI models say? RQ2: What are the relevant guidelines found in RQ1 that can be used in RE activities?

RQ3: What are the advantages and the limitations the identified guidelines provide for the usage of LLMs in RE?

II. Planning the review

The 5 databases used for this review were ACM, Scopus, IEEE Explore, Science Direct, and arXiv. In order to find the primary studies and ensure their relevancy for this SLR, inclusion and exclusion criteria were applied to the search. Following Table shows the criteria used for filtering the papers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023

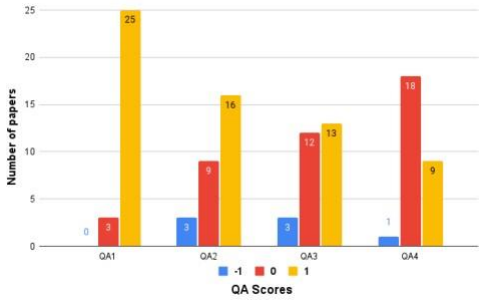
DOI: 10.17148/IJARCCE.2023.121117

INCLUSION AND EXCLUSION CRITERIA FOR THE REVIEW.

Inclusion criteria	Exclusion criteria
Written in English language	Papers with sections or content in languages other than English
Date of publication from 2018	Published prior to 2018
Emphasizes generative AI models	Unrelated to generative AI models
Focus on Natural language prompts	Emphasis on model tuning
Relevant to RQ1	Does not contain PE guidelines

III. Conducting the review

1) Study selection result: After applying the search string to the chosen databases, 271 studies were included for the initial screening. Post the evaluation using inclusion and exclusion criteria and removing duplicates, 28 (10,3%) were recognized as primary studies. The result can be seen in Table above. **2) Criteria for quality assessment:** When the primary studies had been identified, further evaluation through quality assessment was conducted. The obtained graph is



QUALITY ASSESSMENT CRITERIA

Item	Assessment Criteria	Score Scale
QA1	To what extent does the study demonstrate that it has achieved its stated objective(s) in a concrete and detailed way?	1
		0
		-1
QA2	Are the limitations of the study clearly described and discussed?	1
		0
		-1
QA3	Does the study provide contribution to the field of prompt engineering?	1
		0
		-1
QA4	Does the study provide insight to better understand how to prompt large generative AI models?	1
		0
		-1

3) Data extraction: Data collection and synthesis: In order to conduct extraction of relevant data, the form depicted in Table IV was used.

Data	Description	Relevant RQ
DOI	The unique document identifier	General
Year		General
Model type	Text-to-image, text-to-text, GPT-3, BLOOM, Codex, etc.	RQ1
Prompt method	What PE techniques were studied?	RQ1
Guidelines	What guidelines are presented?	RQ1
Findings	Strengths or limitations of the guidelines presented	RQ1

4) The guidelines include: *Context, Template, Persona, Disambiguation, Reasoning, Analysis, Keywords, Wordings, Shorten, Few-shot prompts.*

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023

DOI: 10.17148/IJARCCE.2023.121117

Theme	Description
Context (C)	1. Adding context to examples in prompts produce more efficient and informative output. [51] 2. Provide context to all prompts to avoid output hallucinations. [52] 3. Provide context of the prompt to ensure a closely related output. [53] 4. Use open-ended prompts to generate context before providing the intended question(s). [38] 5. Provide context to the topic of the prompt before describing a task. [54] 6. Adding context tokens to enhance the prompt, improves the related output. [55] 7. The more context tokens pre-appended to prompts, the more fine-grained output.[55]
Persona (P)	1. Improves the generation quality by conditioning the prompt with an identity, such as "Python programmer" or "Math tutor" [56] 2. To explore the requirements of a software-reliant system, include: - "I want you to act as the system", - "Use the requirements to guide your behavior", - "I will ask you to do X, and you will tell me if X is possible given the requirements.", - "If X is possible, explain why using the requirements.", - "If I can't do X based on the requirements, write the missing requirements needed in the format Y." [24]
Templates (T)	1. To improve reasoning and common sense in output, follow a template such as: - "Reason step-by-step for the following problem. [Original prompt inserted here]" [57] 2. The following prompt template has shown an impressive quality of AI art: - "[Medium] [Subject] [Artist(s)] [Details] [Image repository support]" [58]
Disambiguation (D)	1. Ensure any areas of potential miscommunication or ambiguity are caught, by providing a detailed scope: - "Within this scope", - "Consider these requirements or specifications" [24] 2. To find points of weakness in a requirements specification, consider including: - "Point out any areas of ambiguity or potentially unintended outcomes" [24] 3. The persona prompt method can be used to consider potential ambiguities from different perspectives. [22]
Reasoning (R)	1. Prepending "Let's think step by step" improves zero-shot performance. [59] 2. Extending the previously known "Let's think step by step", with "to reach the right conclusion," to highlight decision-making in the prompt. [60] 3. Factual inconsistency evaluation can be significantly boosted using chain-of-thought prompting. [61] 4. Chain Of Thought (CoT) prompting improves LLM performance compared to Zero-shot and without CoT. [62]
Analysis (A)	1. Prepend a prompt in a Zero-shot setting: "Please analyze if the hypothesis is true or false" and use the following template for an analytical output: prompt + approach + premise + hypothesis + "True or False?" [63] 2. ChatGPT models are not "mature enough" for emotional evaluations. [64] 3. Emotion-enhanced CoT prompting is an effective method to leverage emotional cues to enhance the ability of ChatGPT on mental health analysis. [64]
Keywords (K)	1. When picking the prompt, focus on the subject and style keywords instead of connecting words. [34] 2. Pre-appending keywords to prompts are shown to greatly improve performance by providing the language model with appropriate context. [65] 3. Modifiers/Keywords can be added to the details or image repository sections of a template such as: - "[Medium] [Subject] [Artist(s)] [Details] [Image repository support]" [58] 4. The inclusion of multiple descriptive keywords tends to align results closer to expectations. [35]
Wording (W)	1. In translation tasks, adding a newline before the phrase in a new language increases the odds that the output sentence is still English. [51] 2. A complete sentence definition with stop words performs better as a prompt than a set of core terms that were extracted from the complete sentence definition after removing the stop words. [66] 3. Words such as "well-known" and "often used to explain" are successful for analogy generation. [67] 4. Modifying prompts to resemble pseudocode tend to be the most successful in coding tasks. [24], [68] 5. Prompts to contain explicit algorithmic hints in engineering tasks perform better. [68]
Shorten (S)	1. For summarization or text-shortening tasks, the prompt should be written results- and information-oriented, leaving out unnecessary elements. [69]
Few-shot Prompts (F)	1. Inclusion of "Question:" and "Answer:" improves the response, but rarely gives a binary answer. [70] 2. For easier understanding, number examples in few-shot prompting. [71] 3. The format of [INPUT] and [OUTPUT] should linguistically imply the relationship between them. [71] 4. Specifications can be added to each [INPUT] and [OUTPUT] pair to give extra insight into complicated problems. [71] 5. In Few-shot prompting include a rationale in each shot (Input-rationale-output). [72]

5) Mapping through thematic synthesis: After extracting the data from the literature review in order to identify and categorize the PE guidelines, thematic synthesis was conducted with the goal of suggesting a mapping between the guidelines and activities within RE.

RE Activity	Guidelines
Elicitation	C1-C7
Validation	P1-P2
Analysis	R1-R4, T1-T2, A1-A3
Specification	D1-D3
Management	K1-K4

6) Interviews and thematic analysis: In order to further explore the plausibility, as well as limitations and advantages of the suggested mapping, they reached out to 3 experts at different academic institutions, specializing in different areas within RE. The interviews were conducted online over Zoom. The length of the interviews varied slightly but on average lasted around 20 minutes. In order to analyse the interviews which were used to gain perspectives on the suggested mapping, and find answers to RQ3, a thematic analysis was conducted.

As they sought two main points based on RQ3, advantages, and limitations, they also served as two distinctive themes in the analysis. This was possible due to the nature of the interview questions, which asked about the advantages and limitations of the guidelines.

IV. RESULTS

The results of this study are divided into three subsections.

- The first subsection, subsection A, presents their findings, its' details, and answers for RQ1 with a non-exhaustive list of PE guidelines, categorized into the 10 most occurring themes among the identified guidelines.

- In the second subsection, subsection B, they have presented their findings and answer RQ2 by suggesting a mapping of guidelines and themes from RQ1 to various components within RE activities of similar nature.

- The third subsection, subsection C, presents possible advantages and limitations when applying the PE guidelines in RE. By conducting interviews with 3 RE experts, the advantages and limitations of mapped guidelines and RE activities from RQ2 were discussed and established as the answer to RQ3 as well as additional perspectives on RQ2.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 11, November 2023

DOI: 10.17148/IJARCCE.2023.121117

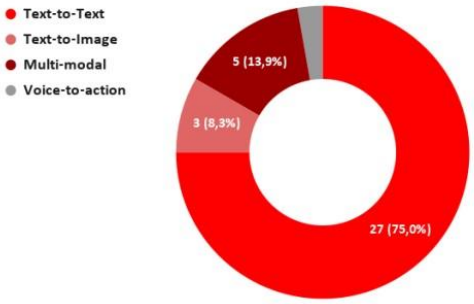


Fig. 2. The number of guidelines per generative model type found in the review

The guidelines include Context, Template, Persona, Disambiguation, Reasoning, Analysis, Keywords, Wordings, Shorten, Few-shot prompts.

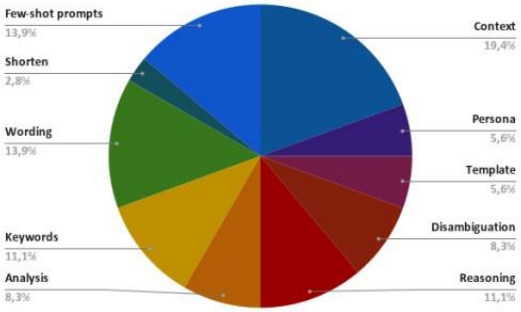


Fig. 3. Theme distribution of the extracted guidelines

Table 1. displays the mapping of the themed guidelines to the respective model type to give a summary of what guidelines originate from which model type. In the same table, they additionally included a mapping of the large generative models that were used in the studies.

Model Type	Models	Guidelines
Text-to-text	GPT-2, GPT-3, GPT-3.5, T0, BLOOM, OPT, InstructGPT, EleutherAI, GPT-J, Galactica, BioBERT Comet, Codex, GitHubCopilot, PaLM-540B	C1, C2, C3, C4, C5, P1, P2, T1 D1, D2, D3, R1, R2, R3, A1, A2, A3, K2 W1, W3, W4, W5 F1, F2, F3, F4, F5
Text-to-image	DALL-E, Midjourney, Imagen, (VQGAN)	T2, K1, K3
Multimodal	CLIP, GPT-4	C6, C7, T1, K4, W2
Voice-to-Action	Undisclosed	S1

Table 1. Model types and models from the review mapped to respective guidelines

V. CONCLUSION

In conclusion, this research paper has provided a comprehensive survey of the advancements in prompt engineering, shedding light on the evolving landscape of natural language processing (NLP) and artificial intelligence (AI). Through an in-depth analysis of various prompt engineering techniques, we have explored how researchers and practitioners are harnessing the power of language models like GPT-3.5 to achieve remarkable results in a myriad of applications.

The survey has highlighted the diverse approaches to prompt engineering, ranging from simple rulebased prompts to more sophisticated methods involving prompt templates, conditioning, and finetuning. The effectiveness of these

DOI: 10.17148/IJARCCE.2023.121117

techniques has been showcased across different domains, including text generation, question-answering, translation, summarization, and more.

Furthermore, the research has underscored the challenges and ethical considerations associated with prompt engineering. Issues such as bias, fairness, and unintended consequences have been discussed, emphasizing the need for responsible AI development. As prompt engineering continues to advance, it is imperative for the AI community to address these challenges and work towards creating models that are not only powerful but also ethical and inclusive.

In conclusion, this survey serves as a valuable resource for researchers, developers, and policymakers seeking to comprehend the current state of prompt engineering and its implications. As we navigate the evolving landscape of AI, this research sets the stage for further exploration and innovation, ultimately contributing to the responsible and impactful development of language models and artificial intelligence technologies.

REFERENCES

[1]. E. A. Olivetti, J. M. Cole, E. Kim, O. Kononova, G. Ceder, T. Y.-J. Han, and A. M. Hiszpanski, Data-driven materials research enabled by natural language processing and information extraction, Applied Physics Reviews 7, 041317 (2020).
[2]. A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, and A. Jain, Structured information extraction from complex scientific text with finetuned large language models
[3]. 10.48550/ARXIV.2212.05238 (2022).
[4]. J. Alammar. Interfaces for explaining transformer language models. https://jalammar.github.io/explainingtransformers/, 2020.
[5]. Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Proceedings of the 2019 Conference of the North American Chapter of the
[6]. Association for Computational Linguistics: Human Language Technologies, Volume 1
[7]. (Long and Short Papers), pages 2357– 2367, Minneapolis, Minnesota.
[8]. Association for Computational Linguistics.
[9]. D. Damian and J. Chisan, “An empirical study of the complex relationships between requirements engineering processes and other processes that lead to payoffs in productivity, quality, and risk management,” IEEE Transactions on Software Engineering, vol. 32, no. 7, pp.
a. 433–453, 2006. DOI:
[10]. 10.1109/TSE.2006.61
[11]. D. Trautmann, A. Petrova, and F. Schilder, “Legal prompt engineering for multilingual legal judgement prediction,” arXiv preprint arXiv:2212.02199, 2022.
[12]. Prompt engineering guide, https : / / www . promptingguide.ai, Accessed: 2023-05-15
[13]. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, Large language models are zero-shot reasoners https://doi.org/10.48550/arXiv.2205.1 1916 (2022).
[14]. OpenAI, Openai prompt design guidelines, https://platform.openai.com/docs/guides/c ompletion/prompt design, Accessed: 202303-10.
[15].] J. F. DeRose, J. Wang, and M. Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. IEEE Transactions on Visualization and Computer Graphics, 27(2):1160–1170, 2020.
[16]. J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, “Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design,” arXiv preprint arXiv:2303.07839, 2023.

