

Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review

Banghao Chen¹, Zhaofeng Zhang¹, Nicolas Langrené^{1*},
Shengxin Zhu^{2,1*}

¹Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Zhuhai 519087, China.

²Research Center for Mathematics, Beijing Normal University, No.18, Jingfeng Road, Zhuhai 519087, Guangdong, China.

*Corresponding author(s). E-mail(s): nicolaslangrene@uic.edu.cn;
Shengxin.Zhu@bnu.edu.cn;
Contributing authors: q030026007@mail.uic.edu.cn;
q030018107@mail.uic.edu.cn;

Abstract

This paper delves into the pivotal role of prompt engineering in unleashing the capabilities of Large Language Models (LLMs). Prompt engineering is the process of structuring input text for LLMs and is a technique integral to optimizing the efficacy of LLMs. This survey elucidates foundational principles of prompt engineering, such as role-prompting, one-shot, and few-shot prompting, as well as more advanced methodologies such as the chain-of-thought and tree-of-thoughts prompting. The paper sheds light on how external assistance in the form of plugins can assist in this task, and reduce machine hallucination by retrieving external knowledge. We subsequently delineate prospective directions in prompt engineering research, emphasizing the need for a deeper understanding of structures and the role of agents in Artificial Intelligence-Generated Content (AIGC) tools. We discuss how to assess the efficacy of prompt methods from different perspectives and using different methods. Finally, we gather information about the application of prompt engineering in such fields as education and programming, showing its transformative potential. This comprehensive survey aims to serve as a friendly guide for anyone venturing through the big world of LLMs and prompt engineering.

Keywords: Prompt engineering, LLM, GPT-4, OpenAI, AIGC, AI agent

1 Introduction

In recent years, a significant milestone in artificial intelligence research has been the progression of natural language processing capabilities, primarily attributed to large language models (LLMs). Many popular models, rooted in the transformer architecture [1], undergo training on extensive datasets derived from web-based text. Central to their design is a self-supervised learning objective, which focuses on predicting subsequent words in incomplete sentences. Those models are called Artificial

Intelligence-Generated Content (AIGC), and their ability to generate coherent and contextually relevant responses is a result of this training process, where they learn to associate words and phrases with their typical contexts.

LLMs operate by encoding the input text into a high-dimensional vector space, where semantic relationships between words and phrases are preserved. The model then decodes this representation to generate a response, guided by the learned statistical patterns [2]. The quality of the response can be influenced by various factors, including the prompt provided to the model, the model’s hyperparameters, and the diversity of the training data.

These models, such as GPT-3 [3], GPT-4 [4], along with many others (e.g., Google’s BARD [5], Anthropic’s Claude2 [6] and Meta’s LLaMA-2 [7]), have been utilized to revolutionize tasks ranging from information extraction to the creation of engaging content [8]. Related to AI systems, the application of LLMs in the workplace has the potential to automate routine tasks, such as data analysis [9] and text generation [10], thereby freeing up time for employees to focus on more complex and rewarding tasks [11]. Furthermore, LLMs have the potential to revolutionize the healthcare sector by assisting in the diagnosis and treatment of diseases. Indeed, by analyzing vast amounts of medical literature, these models can provide doctors with insights into rare conditions, suggest potential treatment pathways, and even predict patient outcomes [12]. In the realm of education, LLMs can serve as advanced tutoring systems, and promote the quality of teaching and learning [13]. Those AI tools can analyze a student’s response, identify areas of improvement, and provide constructive feedback in a coherent and formal manner.

In real applications, the prompt is the input of the model, and its engineering can result in significant output difference [14]. Modifying the structure (e.g., altering length, arrangement of instances) and the content (e.g., phrasing, choice of illustrations, directives) can exert a notable influence on the output generated by the model [15]. Studies show that both the phrasing and the sequence of examples incorporated within a prompt have been observed to exert a substantial influence on the model’s behavior [15, 16].

The discipline of prompt engineering has advanced alongside LLMs. What originated as a fundamental practice of shaping prompts to direct model outputs has matured into a structured research area, replete with its distinct methodologies and established best practices. Prompt engineering refers to the systematic design and optimization of input prompts to guide the responses of LLMs, ensuring accuracy, relevance, and coherence in the generated output. This process is crucial in harnessing the full potential of these models, making them more accessible and applicable across diverse domains. Contemporary prompt engineering encompasses a spectrum of techniques, ranging from foundational approaches such as role-prompting [17] to more sophisticated methods such as “chain of thought” prompting [18]. The domain remains dynamic, with emergent research continually unveiling novel techniques and applications in prompt engineering. The importance of prompt engineering is accentuated by its ability to guide model responses, thereby amplifying the versatility and relevance of LLMs in various sectors. Importantly, a well-constructed prompt can counteract challenges such as machine hallucinations, as highlighted in studies by [19] and [20]. The influence of prompt engineering extends to numerous disciplines. For instance, it has facilitated the creation of robust feature extractors using LLMs, thereby improving their efficacy in tasks such as defect detection and classification [21].

In this paper, we present a comprehensive survey on the prompt engineering of LLMs. The structure of the paper is organized as follows: Section 2 presents the foundational methods of prompt engineering, showcasing various results. It encompasses both basic and advanced techniques. Section 3 further explores advanced methodologies, including the use of external assistance. All the examples are generated on a non-multimodal language generative model, called “default GPT-4”, developed by OpenAI. Section 4 discusses potential future direction in prompt engineering. Section 5 provides insights into prompt evaluation techniques, drawing comparisons between

subjective and objective assessment methods. Finally, Section 6 focuses on the broader applications of prompt engineering across various domains.

2 Basics of prompt engineering

By incorporating just a few key elements, one can craft a basic prompt that enables LLMs to produce high-quality answers. In this section, we discuss some essential components of a well-made prompt.

2.1 Model introduction: GPT-4

All of the output in the following sections are generated by GPT-4, developed by OpenAI [4]. Vast amounts of text data have been used to train GPT-4, whose number of parameters has been estimated to be several orders of magnitude larger than the 175 billion parameters that had been used for the earlier GPT-3 [3]. The architectural foundation of the model rests on transformers [1], which essentially are attention mechanisms that assign varying weights to input data based on the context. Similar to GPT-3, GPT-4 was also fine-tuned to follow a broad class of written instructions by reinforcement learning from human feedback (RLHF) [22, 23], which is a technique that uses human preferences as a reward signal to fine-tune models.

When GPT-4 receives an input prompt, the input text will be firstly converted into tokens that the model can interpret and process. These tokens are then managed by transformer layers, which capture their relationships and context. Within these layers, attention mechanisms distribute different weights to tokens based on their relevance and context. After attention processing, the model forms its internal renditions of the input data, known as intermediate representations. These representations are then decoded back into human-readable text [24].

A significant aspect of this process is the randomness function [25]. This function is influenced by two primary parameters, temperature and top- k sampling. The first one, temperature [26] balances the randomness and determinism in the output. A higher temperature value results in more random outputs, while a lower value makes the output more deterministic. The second one, top- k sampling [27], limits the model's choices to the top k most probable tokens during each step of output generation. The final stage of this process is the output generation, where the model crafts the final text.

2.2 Giving instructions

The method of giving instructions, also considered as re-reading [28], refers to the heuristics of human reading strategy. It has been observed that the output generated by GPT-4, introduced in Section 2.1, tends to be excessively general when provided with fundamental instructions devoid of any supplementary description [29, 30]. An example prompt is shown in Figure 1. When the model is prompted with basic instruction, it faces a plethora of options, which makes the result quite broad. Hence, a comprehensive description is imperative to elicit more precise and relevant outputs [31].

2.3 Be clear and precise

The second basic prompt method is “to be clear and precise”. This involves formulating prompts that are unambiguous and specific, which can guide the model toward generating the desired output.

Most LLM architectures are derived from an extensive array of textual data. It can be conceptualized as a combination of insights from a myriad of authors. When presented with a broad or undetailed prompt, its output predominantly exhibits a generic nature, which, while being applicable across a range of contexts, may not be optimal for any specific application. In contrast, a detailed and precise prompt enables the model to generate content that is more aligned with the unique requirements of

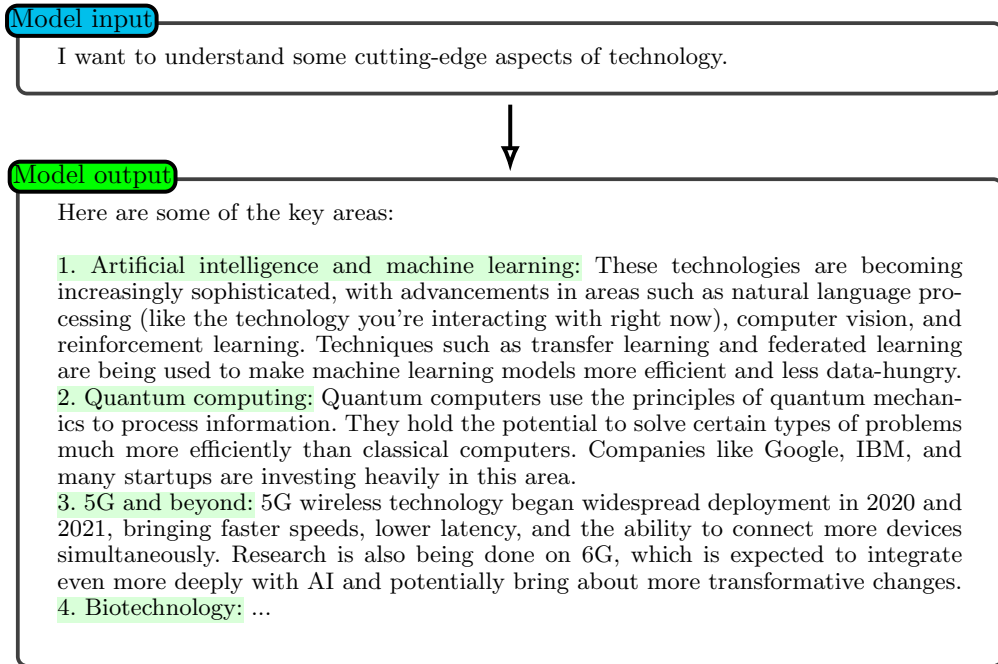


Fig. 1 Giving instructions without extra description.

the given scenario, as it reduces the model's uncertainty and guides it toward the correct response.

For instance, as shown in Figure 2, instead of asking a vague requirement such as “I want to understand the cutting edge of technology.”, a more precise prompt would be “I want to understand the cutting edge of technology, specifically related to artificial intelligence and machine learning...”.

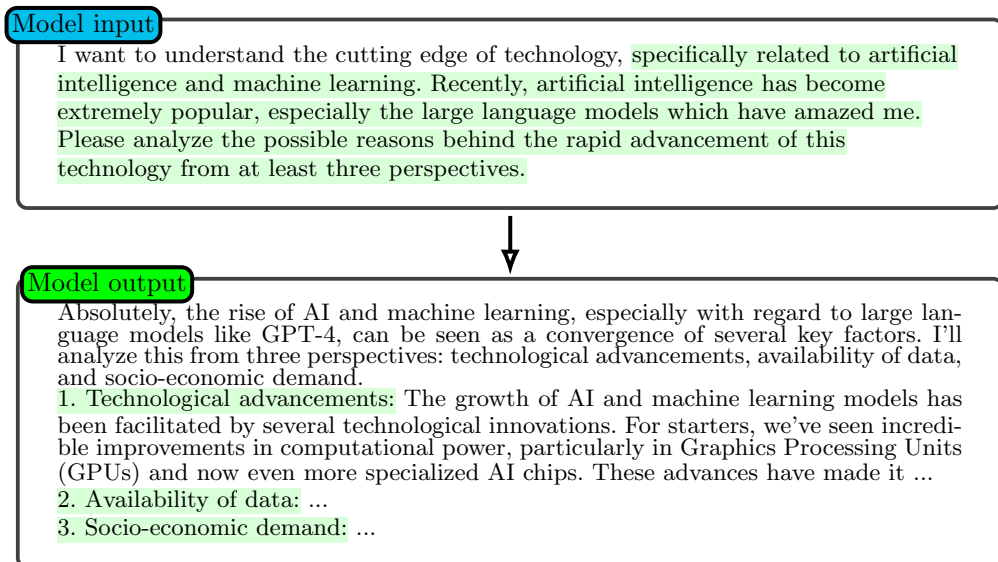


Fig. 2 A clearer and more precise prompt.

2.4 Role-prompting

Role-prompting is another fundamental method in prompt engineering. It involves giving the model a specific role to play, such as a helpful assistant or a knowledgeable expert [32]. This method can be particularly effective in guiding the model's responses

and ensuring that they align with the desired output. For instance, if the model is prompted to act as a historian, it is more likely to provide a detailed and contextually accurate response when asked about a historical event [33]. Another brief example is shown in Figure 3.

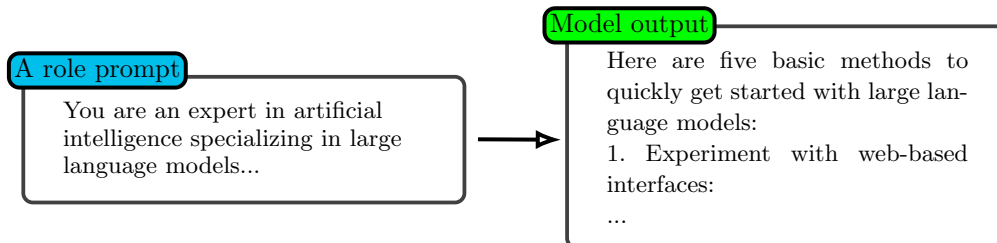


Fig. 3 Role prompting example.

2.5 Use of triple quotes to separate

In prompt engineering, the use of triple quotes is a technique used to separate different parts of a prompt or to encapsulate multi-line strings. This technique is particularly useful when dealing with complex prompts that include multiple components or when the prompt itself contains quotes, which makes the model understand one's instructions better [34].

2.6 Try several times

Due to the non-deterministic nature of LLMs, it is often beneficial to try several times when generating responses. This technique, often referred to as “resampling”, involves running the model multiple times with the same prompt and selecting the best output. This approach can help overcome the inherent variability in the model’s responses and increase the chances of obtaining a high-quality output [26].

2.7 One-shot or few-shot prompting

One-shot and few-shot prompting are two important techniques in prompt engineering. One-shot prompting refers to the method where the model is given a single example to learn from, while few-shot [35] prompting provides the model with multiple examples [36]. The choice between one-shot and few-shot prompting often depends on the complexity of the task and the capability of the model. For instance, for simple tasks or highly capable models, one-shot prompting might be sufficient. An example is shown in Figure 4. However, for more complex tasks or less capable models, few-shot prompting can provide additional context and guidance, thereby improving the model’s performance.

However, as stated in [37], “examples don’t always help”. [37] investigated the intricacies of how large generative language models, such as GPT-3, respond to prompts. One of the significant findings from this paper is that zero-shot prompts can, in certain scenarios, outperform few-shot prompts. This suggests that the role of few-shot examples might not be as much about teaching the model a new task (meta-learning) but rather guiding it to recall a task it has already learned. This insight is crucial as it challenges the conventional wisdom that more examples always lead to better performance [3]. In the context of one-shot or few-shot prompting, it is essential to understand that while examples can guide the model, they do not always enhance its performance. Sometimes, a well-crafted zero-shot prompt can be more effective than providing multiple examples [38].

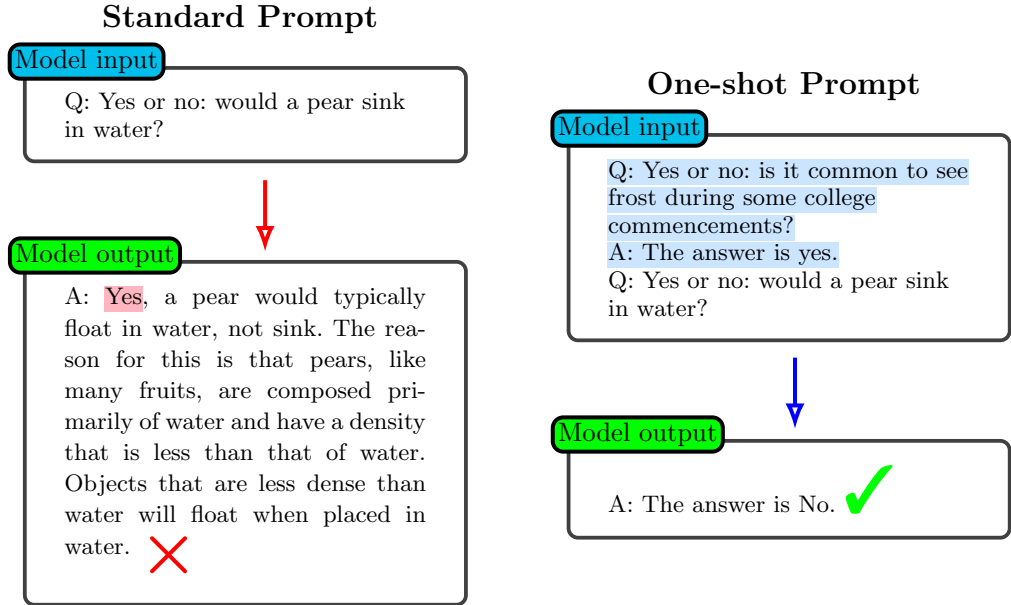


Fig. 4 Comparison of standard prompt and one-shot prompt.

2.8 LLM settings: temperature and top-p

The settings of LLMs, such as the temperature and top- p , play a crucial role in the generation of responses. The temperature parameter controls the randomness of the generated output: a lower temperature leads to more deterministic outputs [39, 40]. The top- p parameter, on the other hand, controls the nucleus sampling [26], which is a method to add randomness to the model’s output [41]. Adjusting these parameters can significantly affect the quality and diversity of the model’s responses, making them essential tools in prompt engineering. However, it has been noted that certain models, exemplified by ChatGPT, do not permit the configuration of these hyperparameters, barring instances where the Application Programming Interface (API) is employed.

3 Advanced methodologies

The foundational methods from the previous section can help us produce satisfactory outputs. However, experiments indicate that when using LLMs for complex tasks such as analysis or reasoning, the accuracy of the model’s outputs still has room for improvement. In this section, we will further introduce advanced techniques in prompt engineering to guide the model in generating more specific, accurate, and high-quality content.

3.1 Chain of thought

The concept of “Chain of Thought” (CoT) prompting [18] in LLMs is a relatively new development in the field of AI, and it has been shown to significantly improve the accuracy of LLMs on various logical reasoning tasks [42–44]. CoT prompting involves providing intermediate reasoning steps to guide the model’s responses, which can be facilitated through simple prompts such as “Let’s think step by step” or through a series of manual demonstrations, each composed of a question and a reasoning chain that leads to an answer [45, 46]. It also provides a clear structure for the model’s reasoning process, making it easier for users to understand how the model arrived at its conclusions.

[47] illustrates the application of CoT prompting to medical reasoning, showing that it can effectively elicit valid intermediate reasoning steps from LLMs. [48] introduces the concept of Self-Education via Chain-of-Thought Reasoning (SECToR), and argues that, in the spirit of reinforcement learning, LLMs can successfully teach

themselves new skills by chain-of-thought reasoning. In another study, [49] used CoT prompting to train verifiers to solve math word problems, demonstrating the technique’s potential in educational applications. [50] proposed a multimodal version of CoT, called Multimodal-CoT, to handle more complex, multimodal tasks beyond simple text-based tasks, such as visual tasks, further expanding the potential applications of CoT.

3.1.1 Zero-shot chain of thought

The concept of “Zero-Shot Chain of Thought” (Zero-shot-CoT) prompting is an advanced iteration of the CoT prompting mechanism, where the “zero-shot” aspect implies that the model is capable of performing some reasoning without having seen any examples of the task during training.

In their research, [51] discovered that the augmentation of queries with the phrase “Let’s think step by step” facilitated the generation of a sequential reasoning chain by LLMs. This reasoning chain subsequently proved instrumental in deriving more precise answers. This technique is based on the idea that the model, much like a human, can benefit from having more detailed and logical steps to process the prompt and generate a response.

For instance, the standard prompt is illustrated in Figure 5, while the appended phrase, “Let’s think step by step”, is depicted in Figure 6. Observations indicate that the incorporation of “Let’s think step by step” enhances the logical coherence and comprehensiveness of the model’s response.

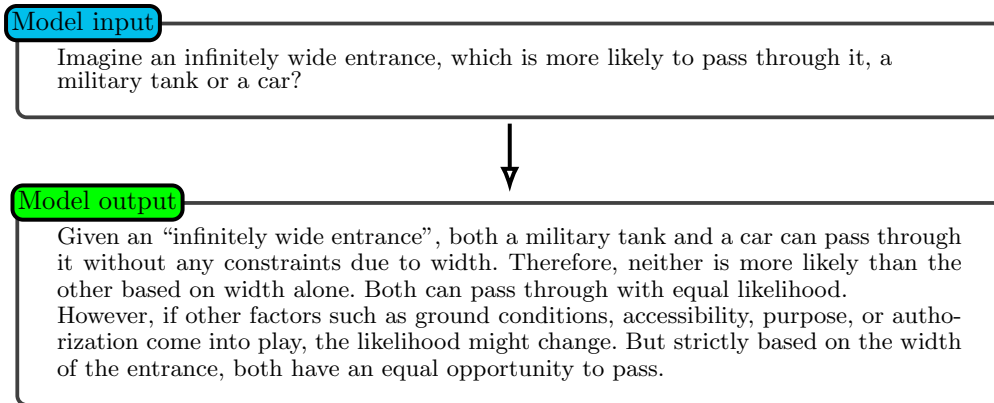


Fig. 5 Standard prompt.

3.1.2 Golden chain of thought

[52] introduced the “golden chain of thought”, providing an innovative approach to generating responses to instruction-based queries. This methodology leverages a set of “ground-truth chain-of-thought” solutions incorporated within the prompt, considerably simplifying the task for the model as it circumvents the necessity for independent CoT generation. Concurrently, a novel benchmark comprising detective puzzles has been designed, to assess the abductive reasoning capacities of LLMs, which is also considered an evaluation of the golden CoT. Finally, according to [52]’s experiment, in the context of the golden CoT, GPT-4 exhibits commendable performance, boasting an 83% solve rate in contrast to the 38% solve rate of the standard CoT.

However, the characteristics of the Golden CoT requiring the “ground-truth chain-of-thought solutions” as an integral part of the prompt also signifies that the Golden CoT’s contribution to solving such problems is limited, despite its high solve rate of 83%.

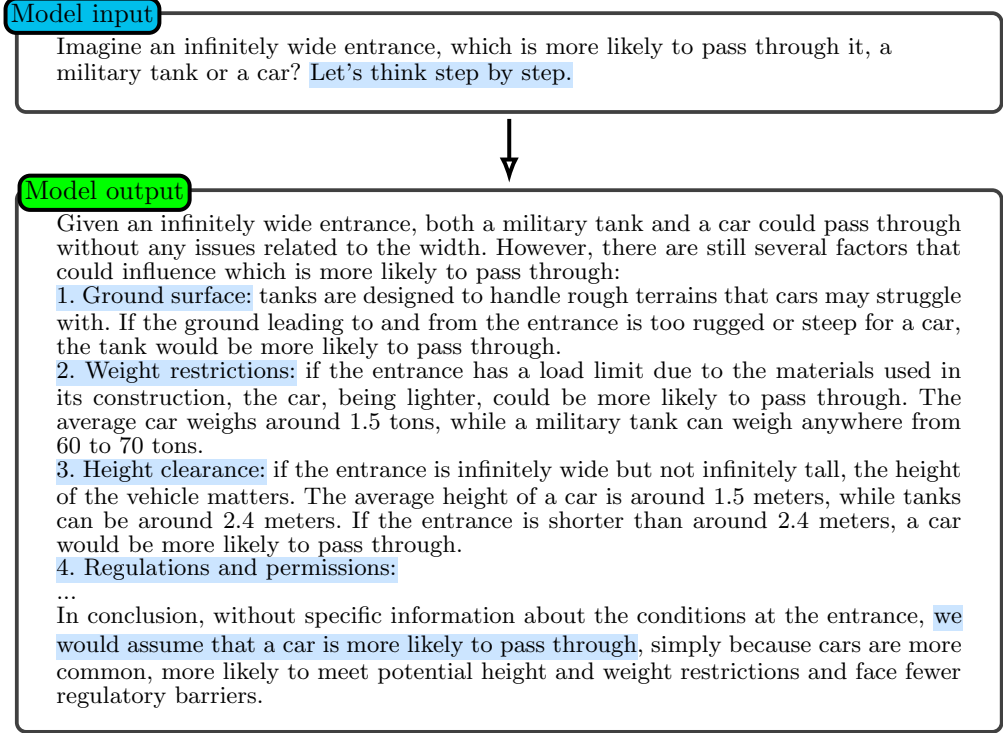


Fig. 6 Adding “Let’s think step by step”.

3.2 Self-consistency

In the assessment of INSTRUCTGPT [53] and GPT-3 [3] on a new synthetic QA dataset called PRONTOQA, for Proof and Ontology-Generated Question-Answering [54, 55], it was observed that although the most extensive model exhibited capability in reasoning tasks, it encountered challenges in proof planning and the selection of the appropriate proof step amidst multiple options, which caused accuracy uncertainties [54]. Self-consistency in LLMs is an advanced prompting technique that aims to ensure the model’s responses are consistent with each other [18, 56], which greatly increases the odds of obtaining highly accurate results. The principle behind it is that if a model is asked to answer a series of related questions, the answers should not contradict each other.

The self-consistency method contains three steps. Firstly, prompt a language model using CoT prompting, then replace the “greedy decode” (1-Best) [25, 57] in CoT prompting by sampling from the language model’s decoder to generate a diverse set of reasoning paths, and finally, marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

It is noteworthy that self-consistency can be harmoniously integrated with most sampling algorithms, including but not limited to, temperature sampling [39, 40], top- k sampling [25, 58, 59], and nucleus sampling [26]. Nevertheless, such an operation may necessitate the invocation of the model’s Application Programming Interface (API) to fine-tune these hyperparameters. In light of this, an alternative approach could be to allow the model to generate results employing diverse reasoning paths, and then generate a diverse set of candidate reasoning paths. The response demonstrating the highest degree of consistency across the various reasoning trajectories is then more inclined to represent the accurate solution [60].

Studies have shown that self-consistency enhances outcomes in arithmetic, commonsense, and symbolic reasoning tasks [2, 61]. Furthermore, in practice, self-consistency can be combined with other techniques to further enhance the model’s performance. For example, a study found that combining self-consistency with a

discriminator-guided multi-step reasoning approach significantly improved the model’s reasoning capabilities [62].

3.3 Generated knowledge

The “generated knowledge” [63] approach in prompt engineering is a technique that leverages the ability of LLMs to generate potentially useful information about a given question or prompt before generating a final response. This method is particularly effective in tasks that require commonsense reasoning, as it allows the model to generate and utilize additional context that may not be explicitly present in the initial prompt.

As exemplified in Figure 5, when posing the query to the model, “Imagine an infinitely wide entrance, which is more likely to pass through it, a military tank or a car?”, standard prompts predominantly yield responses that neglect to factor in the “entrance height”. Conversely, as delineated in Figure 7 and Figure 8, prompting the model to first generate pertinent information and subsequently utilizing generated information in the query leads to outputs with augmented logical coherence and comprehensiveness. Notably, this approach stimulates the model to account for salient factors such as “entrance height”.

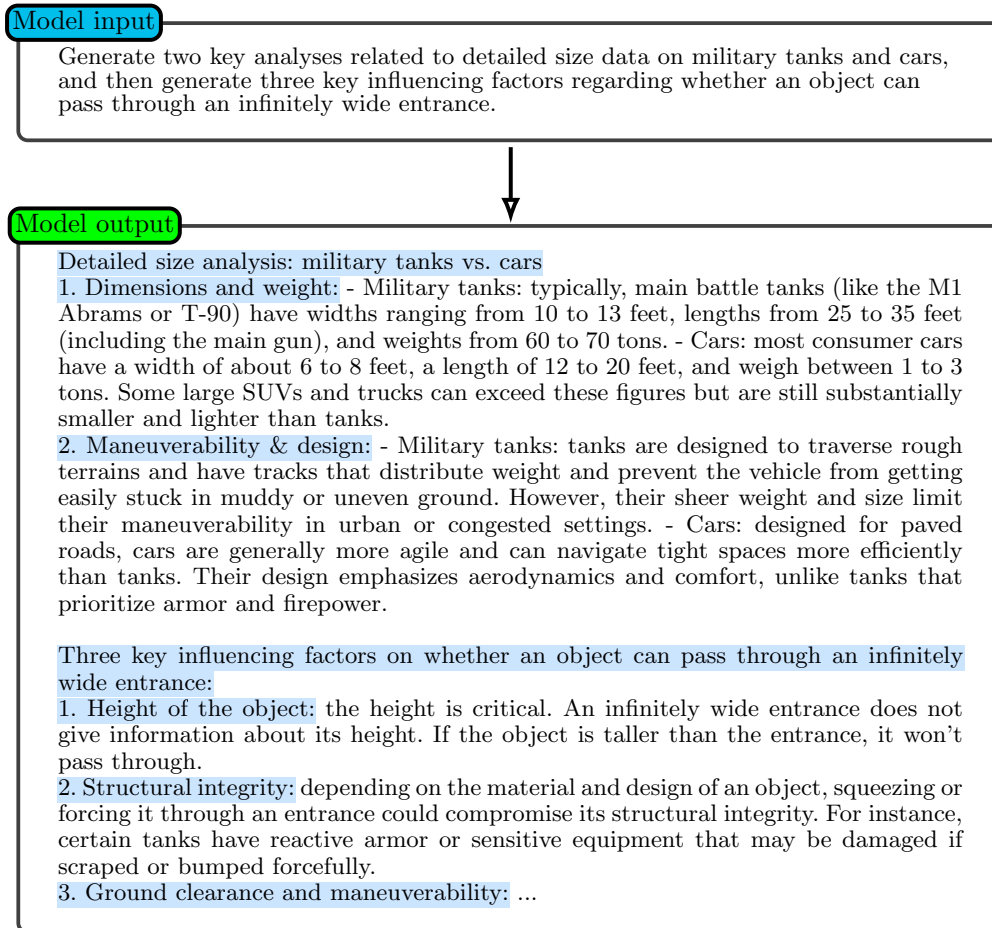


Fig. 7 Generating knowledge (Step1).

3.4 Least-to-most prompting

The concept of “least to most prompting” [64] is an advanced method that involves starting with a minimal prompt and gradually increasing its complexity to elicit more

Model input

“““

Detailed size analysis: military tanks vs. cars

1. Dimensions and weight:

- Military tanks: typically, main battle tanks (like the M1 Abrams or T-90) have widths ranging from 10 to 13 feet, lengths from 25 to 35 feet (including the main gun), and weights from 60 to 70 tons.
- Cars: most consumer cars have a width of about 6 to 8 feet, a length of 12 to 20 feet, and weigh between 1 to 3 tons. Some large SUVs and trucks can exceed these figures but are still substantially smaller and lighter than tanks.

2. Maneuverability & design...

”””

Use the above information to answer the following question: Imagine an infinitely wide entrance, which is more likely to pass through it, a military tank or a car?



Model output

Given the information provided about the dimensions, design, and the factors influencing whether an object can pass through an infinitely wide entrance, we can infer the following:

1. Height of the object: the height is not specified for either tanks or cars. However, given that tanks are built for rough terrains and often have a turret and other equipment on top, they generally have a taller profile compared to most cars. If the entrance had a restrictive height, cars would generally have an advantage here.
2. Structural integrity: ...
3. Ground clearance and maneuverability:

...

Conclusion: If the entrance height is not restrictive and the surface is rough, a military tank would be more likely to pass through it without issues. On the other hand, if the entrance is low-height or has tight squeezes that could compromise the structural integrity of larger, robust vehicles, a car might be at an advantage.

However, given no height constraints and only the factor of width (since it is “infinitely wide”), both a military tank and a car would easily pass through it. In such a scenario, the distinguishing factor would be the terrain of the entrance. If it is rough or uneven, the tank has the advantage. If it is smooth, both would pass with relative ease.

Fig. 8 Combining the generated knowledge with the question (Step2).

sophisticated responses from the language model. The foundational premise of this approach is the decomposition of intricate problems into a succession of more rudimentary subproblems, which are then sequentially addressed. The resolution of each subproblem is expedited by leveraging solutions derived from antecedent subproblems.

Upon rigorous experimentation in domains including symbolic manipulation, compositional generalization, and mathematical reasoning, findings from [64] substantiate that the least-to-most prompting paradigm exhibits the capacity to generalize across challenges of greater complexity than those initially presented in the prompts. They found that LLMs seem to respond effectively to this method, demonstrating its potential for enhancing the reasoning capabilities of these models.

3.5 Tree of thoughts

The “tree of thoughts” (ToT) prompting technique is an advanced method that employs a structured approach to guide LLMs in their reasoning and response generation processes. Unlike traditional prompting methods that rely on a linear sequence of instructions, the ToT method organizes prompts in a hierarchical manner, akin to a tree structure, allowing for deliberate problem-solving [65]. For instance, when tasked with solving a complex mathematical problem, a traditional prompt might directly ask LLMs for the solution. In contrast, using the ToT method, the initial prompt

might first ask the model to outline the steps required to solve the problem. Subsequent prompts would then delve deeper into each step, guiding the model through a systematic problem-solving process.

[65] demonstrates that this formulation is more versatile and can handle challenging tasks where standard prompts might fall short. Another research by [66] further emphasizes the potential of this technique in enhancing the performance of LLMs by structuring their thought processes.

[5] introduces the “tree-of-thought prompting”, an approach that assimilates the foundational principles of the ToT frameworks and transforms them into a streamlined prompting methodology. This technique enables LLMs to assess intermediate cognitive constructs within a singular prompt. An exemplar ToT prompt is delineated in Figure 9 [5].

Tree of thoughts prompting

Imagine three different experts answering this question.
All experts will write down 1 step of their thinking,
then share it with the group.
Then all experts will go on to the next step, etc.
If any expert realizes they’re wrong at any point then they leave.
The question is...

Fig. 9 A sample ToT prompt.

3.6 Graph of thoughts

Unlike the “chain of thoughts” or “tree of thoughts” paradigms, the “graph of thoughts” (GoT) framework [67] offers a more intricate method of representing the information generated by LLMs. The core concept behind GoT is to model this information as an arbitrary graph. In this graph, individual units of information, termed “LLM thoughts”, are represented as vertices. The edges of the graph, on the other hand, depict the dependencies between these vertices. This unique representation allows for the combination of arbitrary LLM thoughts, thereby creating a synergistic effect in the model’s outputs.

In the context of addressing intricate challenges, LLMs utilizing the GoT framework might initially produce several autonomous thoughts or solutions. These individual insights can subsequently be interlinked based on their pertinence and interdependencies, culminating in a detailed graph. This constructed graph permits diverse traversal methods, ensuring the final solution is both precise and comprehensive, encompassing various dimensions of the challenge.

The efficacy of the GoT framework is anchored in its adaptability and the profound insights it can yield, particularly for intricate issues necessitating multifaceted resolutions. Nonetheless, it is imperative to recognize that while GoT facilitates a systematic approach to problem-solving, it also necessitates a profound comprehension of the subject matter and meticulous prompt design to realize optimal outcomes [68].

3.7 Retrieval augmentation

Another direction of prompt engineering is to aim to reduce hallucinations. When using AIGC tools such as GPT-4, it is common to face a problem called “hallucinations”, which refer to the presence of unreal or inaccurate information in the model’s generated output [19, 69]. While these outputs may be grammatically correct, they can be inconsistent with facts or lack real-world data support. Hallucinations arise because the model may not have found sufficient evidence in its training data to support its responses, or it may overly generalize certain patterns when attempting to generate fluent and coherent output [70].

An approach to reduce hallucinations and enhance the effectiveness of prompts is the so-called retrieval augmentation technique, which aims at incorporating up-to-date

external knowledge into the model’s input [71, 72]. It is emerging as an AI framework for retrieving facts from external sources. [73] examines the augmentation of context retrieval through the incorporation of external information. It proposes a sophisticated operation: the direct concatenation of pertinent information obtained from an external source to the prompt, which is subsequently treated as foundational knowledge for input into the expansive language model. Additionally, the paper introduces auto-regressive techniques for both retrieval and decoding, facilitating a more nuanced approach to information retrieval and fusion. This research demonstrates that in-context retrieval-augmented language models [73], when constructed upon readily available general-purpose retrievers, yield significant LLM enhancements across a variety of model dimensions and diverse corpora. In another research, [74] showed that GPT-3 can reduce hallucinations by studying various components of architectures such as Retrieval Augmented Generation (RAG) [75], Fusion-inDecoder (FiD) [76], Seq2seq [77–79] and others. [80] developed Chain-of-Verification (CoVe) to reduce the hallucinations, which introduces that when equipped with tool-use such as retrieval augmentation in the verification execution step, would likely bring further gains.

3.8 Use plugins to polish the prompts

After introducing the detailed techniques and methods of prompt engineering, we now explore the use of some external prompt engineering assistants that have been developed recently and exhibit promising potential. Unlike the methods introduced previously, these instruments can help us to polish the prompt directly. They are adept at analyzing user inputs and subsequently producing pertinent outputs within a context that is defined by itself, thereby amplifying the efficacy of prompts. Some of the plugins provided by OpenAI are good examples of such tools [81].

In certain implementations, the definition of a plugin is incorporated into the prompt, potentially altering the output [82]. Such integration may impact the manner in which LLMs interpret and react to the prompts, illustrating a connection between prompt engineering and plugins. Furthermore, the laborious nature of intricate prompt engineering may be mitigated by plugins, which enable the model to more proficiently comprehend or address user inquiries without necessitating excessively detailed prompts. Consequently, plugins bolster the efficacy of prompt engineering while promoting enhanced user-centric efficiency. These tools, akin to packages, can be seamlessly integrated into Python and invoked directly [83, 84]. Such plugins augment the efficacy of prompts by furnishing responses that are both coherent and contextually pertinent. For instance, the “Prompt Enhancer” plugin [85], developed by AISEO [86], can be invoked by starting the prompt with the word “AISEO” to let the AISEO prompt generator automatically enhance the LLM prompt provided. Similarly, another plugin called “Prompt Perfect”, can be used by starting the prompt with ‘perfect’ to automatically enhance the prompt, aiming for the “perfect” prompt for the task at hand [87, 88].

4 Prospective methodologies

Several key developments on the horizon promise to substantially advance prompt engineering capabilities. In the following section, some of the most significant trajectories would be analyzed that are likely to shape the future of prompt engineering. By anticipating where prompt engineering is headed, developments in this field can be proactively steered toward broadly beneficial outcomes.

4.1 Better understanding of structures

One significant trajectory about the future of prompt engineering that emerges is the importance of better understanding the underlying structures of AI models. This understanding is crucial to effectively guide these models through prompts and to generate outputs that are more closely aligned with user intent.

At the heart of most AI models, including GPT-4, are complex mechanisms designed to understand and generate human language. The interplay of these mechanisms forms the “structure” of these models. Understanding this structure involves unraveling the many layers of neural networks, the various attention mechanisms at work, and the role of individual nodes and weights in the decision-making process of these models [89]. Deepening our understanding of these structures could lead to substantial improvements in prompt engineering. The misunderstanding of the model may cause a lack of reproducibility [90]. By understanding how specific components of the model’s structure influence its outputs, we could design prompts that more effectively exploit these components.

Furthermore, a comprehensive grasp of these structures could shed light on the shortcomings of certain prompts and guide their enhancement. Frequently, the underlying causes for a prompt’s inability to yield the anticipated output are intricately linked to the model’s architecture. For example, [16] found evidence of limitations in previous prompt models and questioned how much these methods truly understood the model.

Exploration of AI model architectures remains a vibrant research domain, with numerous endeavors aimed at comprehending these sophisticated frameworks. A notable instance is DeepMind’s “Causal Transformer” model [91], designed to explicitly delineate causal relationships within data. This represents a stride towards a more profound understanding of AI model architectures, with the potential to help us design more efficient prompts.

Furthermore, a more comprehensive grasp of AI model architectures would also yield advancements in explainable AI. Beyond better prompt engineering, this would also foster greater trust in AI systems and promote their integration across diverse industries [92]. For example, while AI is transforming the financial sector, encompassing areas such as customer service, fraud detection, risk management, credit assessments, and high-frequency trading, several challenges, particularly those related to transparency, are emerging alongside these advancements [93, 94]. Another example is medicine, where AI’s transformative potential faces similar challenges [95, 96].

In conclusion, the trajectory toward a better understanding of AI model structures promises to bring significant advancements in prompt engineering. As we research deeper into these intricate systems, we should be able to craft more effective prompts, understand the reasons behind prompt failures, and enhance the explainability of AI systems. This path holds the potential for transforming how we interact with and utilize AI systems, underscoring its importance in the future of prompt engineering.

4.2 Agent for AIGC tools

The concept of AI agents has emerged as a potential trajectory in AI research [97]. In this brief section, we explore the relationship between agents and prompt engineering and project how agents might influence the future trajectory of AI-generated content (AIGC) tools. By definition, an AI agent comprises large models, memory, active planning, and tool use. AI agents are capable of remembering and understanding a vast array of information, actively planning and strategizing, and effectively using various tools to generate optimal solutions within complex problem spaces [98].

The evolution of AI agents can be delineated into five distinct phases: models, prompt templates, chains, agents, and multi-agents. Each phase carries its specific implications for prompt engineering. Foundational models, exemplified by architectures such as GPT-4, underpin the realm of prompt engineering.

In particular, prompt templates offer an effective way of applying prompt engineering in practice [18]. By using these templates, we can create standardized prompts to guide large models, making the generated output more aligned with the desired outcome. The usage of prompt templates is a crucial step towards enabling AI agents to better understand and execute user instructions.

AI agents amalgamate these methodologies and tools into an adaptive framework. Possessing the capability to autonomously modulate their behaviors and strategies,

they strive to optimize both efficiency and precision in task execution. A salient challenge for prompt engineering emerges: devising and instituting prompts that adeptly steer AI agents toward self-regulation [16].

In conclusion, the introduction of agent-based paradigms heralds a novel trajectory for the evolution of AIGC tools. This shift necessitates a reevaluation of established practices in prompt engineering and ushers in fresh challenges associated with the design, implementation, and refinement of prompts.

5 Assessing the efficacy of prompt methods

There are many different ways to evaluate the quality of the output. To assess the efficacy of current prompt methods in AIGC tools, evaluation methods can generally be divided into subjective and objective categories.

5.1 Subjective and objective evaluations

Subjective evaluations primarily rely on human evaluators to assess the quality of the generated content. Human evaluators can read the text generated by LLMs and score it for quality. Subjective evaluations typically include aspects such as fluency, accuracy, novelty, and relevance [26]. However, these evaluation methods are, by definition, subjective and can be prone to inconsistencies.

Objective evaluations, also known as automatic evaluation methods, use machine learning algorithms to score the quality of text generated by LLMs. Objective evaluations employ automated metrics, such as BiLingual Evaluation Understudy (BLEU) [99], which assigns a score to system-generated outputs, offering a convenient and rapid way to compare various systems and monitor their advancements. Other evaluations such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [100], and Metric for Evaluation of Translation with Explicit ORdering (METEOR) [101], assess the similarity between the generated text and reference text. More recent evaluation methods, such as BERTScore [102], aim to assess at a higher semantic level. However, these automated metrics often fail to fully capture the assessment results of human evaluators and therefore must be used with caution [103].

Subjective evaluation and objective evaluation methods each have their own advantages and disadvantages. Subjective evaluation tends to be more reliable than objective evaluation, but it is also more expensive and time-consuming. Objective evaluation is less expensive and quicker than subjective evaluation. For instance, despite numerous pieces of research highlighting the limited correlation between BLEU and alternative metrics based on human assessments, their popularity has remained unaltered [104, 105]. Ultimately, the best way to evaluate the quality of LLM output depends on the specific application [106]. If quality is the most important factor, then using human evaluators is the better choice. If cost and time are the most important factors, then using automatic evaluation methods is better.

5.2 Comparing different prompt methods

In the field of prompt engineering, previous work has mostly focused on designing and optimizing specific prompting methods, but evaluating and comparing different prompting approaches in a systematic manner remains limited. There are some models that are increasingly used to grade the output of other models, which aim to ‘check’ the ability of other models [107, 108]. For instance, LLM-Eval [109] was developed to measure the open-domain conversations with LLMs. This method tries to evaluate the performance of LLMs on various benchmark datasets [110, 111] and demonstrate their efficiency. Other studies experiment mainly on certain models or tasks and employ disparate evaluation metrics, restricting comparability across methods [112, 113]. Nevertheless, recent research proposed a general evaluation framework called InstructEval [114] that enables a comprehensive assessment of prompting techniques across multiple models and tasks. The InstructEval study reached the following conclusions: in few-shot settings, omitting prompts or using generic task-agnostic prompts tends

to outperform other methods, with prompts having little impact on performance; in zero-shot settings, expert-written task-specific prompts can significantly boost performance, with automated prompts not outperforming simple baselines; the performance of automated prompt generation methods is inconsistent, varying across different models and task types, displaying a lack of generalization. InstructEval provides important references for prompt engineering and demonstrates the need for more universal and reliable evaluation paradigms to design optimal prompts.

6 Applications improved by prompt engineering

The output enhancements provided by prompt engineering techniques make LLMs better applicable to real-world applications. This section briefly discusses applications of prompt engineering in fields such as teaching, programming, and others.

<p>Principles of Accounting: Course Outline Duration: 10 weeks Class Structure: <ul style="list-style-type: none"> •Lecture: 1 hour per week •Workshop/Tutorial: 2 hours per week Week 1: Introduction to Accounting <ul style="list-style-type: none"> •Lecture: The Importance of Accounting in Business •Workshop: Introduction to Financial Statements Week 2: Basics of Bookkeeping <ul style="list-style-type: none"> •Lecture: The Accounting Cycle •Workshop: Recording Transactions using Double Entry System </p>	<p>Group Assignment: Company Financial Analysis •Description: Groups of 4 students will select a public company and perform a financial statement analysis ... •Rubrics: <ul style="list-style-type: none"> • Research and Data Collection (20%) <ul style="list-style-type: none"> • Correct annual report selected • All relevant data extracted properly • Analysis (40%) <ul style="list-style-type: none"> • ... • Presentation (20%) <ul style="list-style-type: none"> • ... • </p>
--	--

Fig. 10 Guideline of courses generated by GPT-4

6.1 Assessment in teaching and learning

The study [115] investigates the application of machine learning methods in young student education. In such a context, prompt engineering can facilitate the creation of personalized learning environments. By offering tailored prompts, LLMs can adapt to an individual’s learning pace and style. Such an approach can allow for personalized assessments and educational content, paving the way for a more individual-centric teaching model. Recent advancements in prompt engineering suggest that AI tools can also cater to students with specific learning needs, thus fostering inclusivity in education [116]. As a simple example, it is possible for professors to provide rubrics or guidelines for a future course with the assistance of AI. As Figure 10 shows, when GPT-4 was required to provide a rubric about a course, with a suitable prompt, it was able to respond with a specific result that may satisfy the requirement.

The advancements in prompt engineering also bring better potential for automated grading in education. With the help of sophisticated prompts, LLMs can provide preliminary assessments, reducing the workload for educators while providing instant feedback to students [117]. Similarly, these models, when coupled with well-designed prompts, can analyze a vast amount of assessment data, thus providing valuable insights into learning patterns and informing educators about areas that require attention or improvement [118, 119].

6.2 Content creation and editing

With controllable improved input, LLMs have primarily been used in creative works, such as content creation. Pathways Language Model (PaLM) [57] and prompting approach have been used to facilitate cross-lingual short story generation [10]. The Recursive Reprompting and Revision framework (Re³) [120] employs zero-shot

prompting [51] with GPT-3 to craft a foundational plan including elements such as settings, characters, and outlines. Subsequently, it adopts a recursive technique, dynamically prompting GPT-3 to produce extended story continuations. For another example, Detailed Outline Control (DOC) [121] aims at preserving plot coherence across extensive texts generated with the assistance of GPT-3. Unlike Re³, DOC employs a detailed outliner and detailed controller for implementation. The detailed outliner initially dissects the overarching outline into subsections through a breadth-first method, where candidate generations for these subsections are generated, filtered, and subsequently ranked. This process is similar to the method of chain-of-thought (in Section 3.1). Throughout this generation process, an OPT-based Future Discriminators for Generation (FUDGE) [122] detailed controller plays a crucial role in maintaining relevance.

6.3 Computer programming

Prompt engineering can help LLMs perform better at outputting programming codes. By using a self-debugging prompting approach [46], which contains simple feedback, unit-test, and code explanation prompts module, the text-to-SQL [123] model is able to provide a solution it can state as correct unless the maximum number of attempts has been reached. Another example, Multi-Turn Programming Benchmark (MTPB) [124], was constructed to implement a program by breaking it into multi-step natural language prompts.

Another approach is provided in [125], which introduced the Repo-Level Prompt Generator (RLPG) to dynamically retrieve relevant repository context and construct a prompt for a given task, especially on code auto-completion task. The most suitable prompt is selected by a prompt proposal classifier and combined with the default context to generate the final output.

6.4 Reasoning tasks

AIGC tools have shown promising performance in reasoning tasks. Several previous researches found that few-shot prompting can enhance the performance in generating accurate reasoning steps for word-based math problems in the GSM8K dataset [44, 49, 56, 57]. The strategy of including the reasoning traces in such as few-shot prompts [35], self-talk [126] and chain-of-thought [18], was shown to encourage the model to generate verbalized reasoning steps. [127] conducted experiments by involving prompting strategies, various fine-tuning techniques, and re-ranking methods to assess their impact on enhancing the performance of a base LLM. They found that a customized prompt significantly improved the model’s ability with fine-tuning, and demonstrated a significant advantage by generating substantially fewer errors in reasoning. In another research, [51] observed that solely using zero-shot CoT prompting leads to a significant enhancement in the performance of GPT-3 and PaLM when compared to the conventional zero-shot and few-shot prompting methods. This improvement is particularly noticeable when evaluating these models on the MultiArith [128] and GSM8K [49] datasets. [129] also introduced a novel prompting approach called Diverse Verifier on Reasoning Step (DIVERSE). This approach involves using a diverse set of prompts for each question and incorporates a trained verifier with an awareness of reasoning steps. The primary aim of DIVERSE is to enhance the performance of GPT-3 on various reasoning benchmarks, including GSM8K and others. All these works show that in the application of reasoning tasks, properly customized prompts can obtain better results from the model.

6.5 Dataset generation

LLMs possess the capability of in-context learning, enabling them to be effectively prompted to generate synthetic datasets for training smaller, domain-specific models. [130] put forth three distinct prompting approaches for training data generation using GPT-3: unlabeled data annotation, training data generation, and assisted training

data generation. Besides, [131] is designed for the generation of supplementary synthetic data for classification tasks. GPT-3 is utilized in conjunction with a prompt that includes real examples from an existing dataset, along with a task specification. The goal is to jointly create synthetic examples and pseudo-labels using this combination of inputs.

7 Conclusion

In this paper, we present a comprehensive overview of prompt engineering techniques and their instrumental role in refining Large Language Models (LLMs). We detail both foundational and advanced methodologies in prompt engineering, illustrating their efficacy in directing LLMs toward targeted outputs. We also analyze retrieval augmentation and plugins, which can further augment prompt engineering. We discuss broader applications of prompt engineering, highlighting its potential in sectors such as education and programming. We finally cast a forward-looking gaze on the future avenues of prompt engineering, underscoring the need for a deeper understanding of LLM architectures and the significance of agent-based paradigms. In summary, prompt engineering has emerged as a critical technique for guiding and optimizing LLMs. As the ubiquity of prompt engineering develops, we hope that this paper can lay the groundwork for further research.

8 Acknowledgement

This work was funded by the Natural Science Foundation of China (12271047); Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College (2022B1212010006); UIC research grant (R0400001-22; UICR0400008-21; UICR0700041-22; R72021114); Guangdong College Enhancement and Innovation Program (2021ZDZX1046).

References

- [1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17; 2017. p. 6000–6010.
- [2] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021. p. 610–623.
- [3] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. language models Are Few-Shot Learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS’20; 2020. .
- [4] OpenAI. GPT-4 Technical Report; 2023. ArXiv:2303.08774.
- [5] Hulbert D.: Tree of knowledge: ToK aka Tree of Knowledge dataset for Large Language Models LLM. Accessed: 2023-8-15. figshare <https://github.com/dave1010/tree-of-thought-prompting>.
- [6] Anthropic.: Claude2. Accessed: 2023-7-11. figshare <https://www.anthropic.com>.
- [7] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models; 2023. ArXiv:2307.09288.
- [8] Sarkhel R, Huang B, Lockard C, Shiralkar P. Self-training for label-efficient information extraction from semi-structured web-pages. Proceedings of the VLDB Endowment. 2023;16(11):3098–3110.

- [9] Cheng L, Li X, Bing L. Is GPT-4 a good data analyst?; 2023. ArXiv:2305.15038.
- [10] Razumovskaia E, Maynez J, Louis A, Lapata M, Narayan S. Little red riding hood goes around the globe: crosslingual story planning and generation with large language models; 2022. ArXiv:2212.10471.
- [11] Manning S, Mishkin P, Hadfield G, Eloundou T, Eisner E. A research agenda for assessing the economic impacts of code generation models; 2022.
- [12] Choudhury A, Asan O, et al. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR medical informatics*. 2020;8(7):e18599.
- [13] Baidoo-Anu D, Ansah LO. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*. 2023;7(1):52–62.
- [14] Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models; 2023. ArXiv:2307.10169.
- [15] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*; 2022. p. 8086–8098.
- [16] Webson A, Pavlick E. Do prompt-based models really understand the meaning of their prompts? In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2022. p. 2300–2344.
- [17] Shanahan M, McDonell K, Reynolds L. Role-play with large language models; 2023. ArXiv:2305.16367.
- [18] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: *Advances in Neural Information Processing Systems*. vol. 35; 2022. p. 24824–24837.
- [19] Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020. p. 1906–1919.
- [20] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4; 2023. ArXiv:2303.12712.
- [21] Yong G, Jeon K, Gil D, Lee G. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Computer-Aided Civil and Infrastructure Engineering*. 2022;38(11):1536–1554.
- [22] Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*. 2017;30.
- [23] Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, et al. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*. 2020;33:3008–3021.
- [24] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training; 2018.

<https://openai.com/research/language-unsupervised>.

- [25] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al.: Language models are unsupervised multitask learners. Assessed: 2019-02-07. figshare https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- [26] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. In: International Conference on Learning Representations; 2020. .
- [27] Welleck S, Kulikov I, Roller S, Dinan E, Cho K, Weston J. Neural text generation with unlikelihood training; 2019. ArXiv:1908.04319.
- [28] Xu X, Tao C, Shen T, Xu C, Xu H, Long G, et al. Re-reading improves reasoning in language models; 2023. ArXiv:2309.06275.
- [29] YanSong S, JingLi Tencent A. Joint learning embeddings for Chinese words and their components via ladder structured networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18); 2018. p. 4375–4381.
- [30] Luo L, Ao X, Song Y, Li J, Yang X, He Q, et al. Unsupervised neural aspect extraction with sememes. In: IJCAI; 2019. p. 5123–5129.
- [31] Yang M, Qu Q, Tu W, Shen Y, Zhao Z, Chen X. Exploring human-like reading strategy for abstractive text summarization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33; 2019. p. 7362–7369.
- [32] Zhang Z, Gao J, Dhaliwal RS, Jia-Jun Li T. VISAR: a human-AI argumentative writing assistant with visual programming and rapid draft prototyping; 2023. ArXiv:2304.07810.
- [33] Buren DV. Guided scenarios with simulated expert personae: a remarkable strategy to perform cognitive work; 2023. ArXiv:2306.03104.
- [34] OpenAI.: Tactic: use delimiters to clearly indicate distinct parts of the input. Accessed: 2023-09-01. figshare <https://platform.openai.com/docs/guides/gpt-best-practices/tactic-use-delimiters-to-clearly-indicate-distinct-parts-of-the-input>.
- [35] Logan IV R, Balažević I, Wallace E, Petroni F, Singh S, Riedel S. Cutting down on prompts and parameters: simple few-shot learning with language models. In: Findings of the Association for Computational Linguistics: ACL 2022; 2022. p. 2824–2835.
- [36] Shyr C, Hu Y, Harris PA, Xu H. Identifying and extracting rare disease phenotypes with Large language models; 2023. ArXiv:2306.12656.
- [37] Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–7.
- [38] Liu J, Gardner M, Cohen SB, Lapata M. Multi-step inference for reasoning over paragraphs; 2020. ArXiv:2004.02995.
- [39] Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for Boltzmann machines. *Cognitive Science*. 1985;9(1):147–169.
- [40] Fidler J, Goldberg Y. Controlling linguistic style aspects in neural language generation. In: Proceedings of the Workshop on Stylistic Variation; 2017. p.

- [41] Xu C, Guo D, Duan N, McAuley J. RIGA at SemEval-2023 Task 2: NER enhanced with GPT-3. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*; 2023. p. 331–339.
- [42] Wu S, Shen EM, Badrinath C, Ma J, Lakkaraju H. Analyzing chain-of-thought prompting in Large language models via gradient-based feature Attributions; 2023. ArXiv:2307.13339.
- [43] Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in Large language models. In: *Eleventh International Conference on Learning Representations*; 2023. .
- [44] Lewkowycz A, Andreassen A, Dohan D, Dyer E, Michalewski H, Ramasesh V, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*. 2022;35:3843–3857.
- [45] Zhou H, Nova A, Larochelle H, Courville A, Neyshabur B, Sedghi H. Teaching Algorithmic Reasoning via In-context Learning; 2022. ArXiv:2211.09066.
- [46] Lee N, Sreenivasan K, Lee JD, Lee K, Papailiopoulos D. Teaching arithmetic to small transformers; 2023. ArXiv:2307.03381.
- [47] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Large language models perform diagnostic reasoning. In: *Eleventh International Conference on Learning Representations*; 2022. .
- [48] Zhang H, Parkes DC. Chain-of-thought reasoning is a policy improvement operator; 2023. ArXiv:2309.08589.
- [49] Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, et al. Training verifiers to solve math word problems; 2021. ArXiv:2110.14168.
- [50] Huang S, Dong L, Wang W, Hao Y, Singhal S, Ma S, et al. Language is not all you need: aligning perception with language models; 2023. ArXiv:2302.14045.
- [51] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*. 2022;35:22199–22213.
- [52] Del M, Fishel M. True detective: a deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*; 2023. .
- [53] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*. 2022;35:27730–27744.
- [54] Saparov A, He H. Language models are greedy reasoners: a systematic formal analysis of chain-of-thought; 2022. ArXiv:2210.01240.
- [55] Tafjord O, Dalvi B, Clark P. ProofWriter: generating implications, proofs, and abductive statements over natural language. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; 2021. p. 3621–3634.
- [56] Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. In: *Eleventh International Conference on Learning Representations*; 2023. .

- [57] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways; 2022. ArXiv:2204.02311.
- [58] Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018. p. 889–898.
- [59] Holtzman A, Buys J, Forbes M, Bosselut A, Golub D, Choi Y. Learning to write with cooperative discriminators. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018. p. 1638–1649.
- [60] Huang J, Gu SS, Hou L, Wu Y, Wang X, Yu H, et al. Large language models can self-improve; 2022. ArXiv:2210.11610.
- [61] Shum K, Diao S, Zhang T. Automatic prompt augmentation and selection with chain-of-thought from labeled data; 2023. ArXiv:2302.12822.
- [62] Khalifa M, Logeswaran L, Lee M, Lee H, Wang L. Discriminator-guided multi-step reasoning with language models; 2023. ArXiv:2305.14934.
- [63] Liu J, Liu A, Lu X, Welleck S, West P, Le Bras R, et al. Generated knowledge prompting for commonsense reasoning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 3154–3169.
- [64] Zhou D, Schärli N, Hou L, Wei J, Scales N, Wang X, et al. Least-to-most prompting enables complex reasoning in Large language models. In: Eleventh International Conference on Learning Representations; 2023. .
- [65] Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models; 2023. ArXiv:2305.10601.
- [66] Long J. Large language model guided tree-of-thought; 2023. ArXiv:2305.08291.
- [67] Besta M, Blach N, Kubicek A, Gerstenberger R, Gianinazzi L, Gajda J, et al. Graph of thoughts: solving elaborate problems with large language models; 2023. ArXiv:2308.09687.
- [68] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents; 2023. ArXiv:2308.11432.
- [69] Lee K, Firat O, Agarwal A, Fannjiang C, Sussillo D. Hallucinations in neural machine translation; 2018.
- [70] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*. 2023;55(12):1–38.
- [71] Lazaridou A, Gribovskaya E, Stokowiec W, Grigorev N. Internet-augmented language models through few-shot prompting for open-domain question answering; 2022. ArXiv:2203.05115.
- [72] Jiang Z, Xu FF, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, et al. Active retrieval augmented generation; 2023. ArXiv:2305.06983.
- [73] Ram O, Levine Y, Dalmedigos I, Muhlgay D, Shashua A, Leyton-Brown K, et al. In-context retrieval-augmented language models; 2023. ArXiv:2302.00083.
- [74] Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation; 2021. ArXiv:2104.07567.

- [75] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*. 2020;33:9459–9474.
- [76] Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering; 2020. ArXiv:2007.01282.
- [77] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020. p. 7871–7880.
- [78] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. 2020;21(1):5485–5551.
- [79] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, et al. Recipes for building an open-domain chatbot. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; 2021. p. 300–325.
- [80] Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, et al. Chain-of-verification reduces hallucination in large language models; 2023. ArXiv:2309.11495.
- [81] OpenAI.: ChatGPT plugins. Accessed: 2023-10-15. figshare <https://openai.com/blog/chatgpt-plugins>.
- [82] Bisson S.: Microsoft build 2023: Microsoft extends its copilots with open standard plugins. Accessed: 2023-05-25. figshare <https://www.techrepublic.com/article/microsoft-extends-copilot-with-open-standard-plugins/>.
- [83] Ng A.: ChatGPT prompt engineering for developers. Accessed: 2023-07-18. figshare <https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>.
- [84] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, et al. Recipes for building an open-domain chatbot. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; 2021. p. 300–325.
- [85] whatplugin.: Prompt enhancer & ChatGPT plugins for AI development tools like prompt enhancer. Accessed: 2023-09-14. figshare <https://www.whatplugin.ai/plugins/prompt-enhancer>.
- [86] : AISEO. Accessed: 2023-8-15. figshare <https://aiseo.ai/>.
- [87] for Search Engines C.: Prompt perfect plugin for ChatGPT. Accessed: 2023-10-15. figshare <https://chatonai.org/prompt-perfect-chatgpt-plugin>.
- [88] Prompt Perfect.: Terms of service. Accessed: 2023-09-20. figshare <https://promptperfect.xyz/static/terms.html>.
- [89] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*. 2020;23(1):18.
- [90] Recht B, Re C, Wright S, Niu F. Hogwild!: a lock-free approach to parallelizing stochastic gradient descent. In: *Advances in neural information processing systems*. vol. 24; 2011. .

- [91] Melnychuk V, Frauen D, Feuerriegel S. Causal transformer for estimating counterfactual outcomes. In: International Conference on Machine Learning; 2022. p. 15293–15329.
- [92] Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*. 2023;24(2):125–137.
- [93] Bertucci L, Brière M, Fliche O, Mikael J, Szpruch L. Deep learning in finance: from implementation to regulation; 2022. SSRN:4080171.
- [94] Maple C, Szpruch L, Epiphanion G, Staykova K, Singh S, Penwarden W, et al. The AI revolution: opportunities and challenges for the finance sector; 2023. ArXiv:2308.16538.
- [95] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*. 2020;20(1):1–9.
- [96] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature Medicine*. 2022;28(1):31–38.
- [97] Öztürk D. What does artificial intelligence mean for organizations? A systematic review of organization studies research and a way forward. *The Impact of Artificial Intelligence on Governance, Economics and Finance, Volume I*. 2021;p. 265–289.
- [98] Seeamber R, Badea C. If our aim is to build morality into an artificial agent, how might we begin to go about doing so? *IEEE Intelligent Systems*. 2023;.
- [99] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*; 2002. p. 311–318.
- [100] Chin-Yew L. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*; 2004. p. 74–81.
- [101] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; 2005. p. 65–72.
- [102] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: *International Conference on Learning Representations*; 2020. .
- [103] Sai AB, Mohankumar AK, Khapra MM. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*. 2022;55(2):1–39.
- [104] Ananthakrishnan R, Bhattacharyya P, Sasikumar M, Shah RM. Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *Icon*. 2007;64.
- [105] Callison-Burch C, Osborne M, Koehn P. Re-evaluating the role of BLEU in machine translation research. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*; 2006. p. 249–256.
- [106] Stent A, Marge M, Singhai M. Evaluating evaluation methods for generation in the presence of variation. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer; 2005. p. 341–351.

- [107] Jain N, Saifullah K, Wen Y, Kirchenbauer J, Shu M, Saha A, et al. Bring your own data! Self-supervised evaluation for large language models; 2023. ArXiv:2306.13651.
- [108] Wang Y, Yu Z, Zeng Z, Yang L, Wang C, Chen H, et al. PandaLM: an automatic evaluation benchmark for LLM instruction tuning optimization; 2023. ArXiv:2306.05087.
- [109] Lin YT, Chen YN. LLM-eval: unified multi-dimensional automatic evaluation for open-domain conversations with large language models; 2023. ArXiv:2305.13711.
- [110] Kiela D, Bartolo M, Nie Y, Kaushik D, Geiger A, Wu Z, et al. Dynabench: rethinking benchmarking in NLP; 2021. ArXiv:2104.14337.
- [111] Dehghani M, Tay Y, Gritsenko AA, Zhao Z, Houlsby N, Diaz F, et al. The benchmark lottery; 2021. ArXiv:2107.07002.
- [112] Deng M, Wang J, Hsieh CP, Wang Y, Guo H, Shu T, et al. RLPrompt: optimizing discrete text prompts with reinforcement learning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022. p. 3369–3391.
- [113] Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, et al. Large language models are human-level prompt engineers. In: Eleventh International Conference on Learning Representations; 2022. .
- [114] Ajith A, Pan C, Xia M, Deshpande A, Narasimhan K. InstructEval: systematic evaluation of instruction selection methods; 2023. ArXiv:2307.00259.
- [115] Tang J, Zhou X, Wan X, Daley M, Bai Z. ML4STEM professional development program: enriching K-12 STEM teaching with machine learning. *International Journal of Artificial Intelligence in Education*. 2023;33(1):185–224.
- [116] Xie Q, Dai Z, Hovy E, Luong MT, Le QV. Unsupervised data augmentation for consistency training. In: Advances in neural information processing systems. vol. 33; 2020. p. 6256–6268.
- [117] Ariely M, Nazaretsky T, Alexandron G. Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*. 2023;33(1):1–34.
- [118] Nilsson F, Tuvstedt J. GPT-4 as an automatic grader: the accuracy of grades set by GPT-4 on introductory programming assignments [Bachelor Thesis]. KTH Royal Institute of Technology; 2023.
- [119] Schneider J, Richner R, Riser M. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*. 2023;33(1):88–118.
- [120] Yang K, Peng N, Tian Y, Klein D. Re3: generating longer stories with recursive reprompting and revision; 2022. ArXiv:2210.06774.
- [121] Yang K, Klein D, Peng N, Tian Y. Doc: improving long story coherence with detailed outline control; 2022. ArXiv:2212.10077.
- [122] Yang K, Klein D. FUDGE: controlled text generation with future discriminators; 2021. ArXiv:2104.05218.
- [123] Elgohary A, Hosseini S, Awadallah AH. Speak to your parser: interactive text-to-SQL with natural language feedback. In: Annual Meeting of the Association

for Computational Linguistics; 2020. .

- [124] Nijkamp E, Pang B, Hayashi H, Tu L, Wang H, Zhou Y, et al. Codegen: an open large language model for code with multi-turn program synthesis; 2022. ArXiv:2203.13474.
- [125] Shrivastava D, Larochelle H, Tarlow D. Repository-level prompt generation for large language models of code. In: International Conference on Machine Learning; 2023. p. 31693–31715.
- [126] Shwartz V, West P, Bras RL, Bhagavatula C, Choi Y. Unsupervised common-sense question answering with self-talk; 2020. ArXiv:2004.05483.
- [127] Uesato J, Kushman N, Kumar R, Song F, Siegel N, Wang L, et al. Solving math word problems with process-and outcome-based feedback; 2022. ArXiv:2211.14275.
- [128] Roy S, Roth D. Solving general arithmetic word problems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015. p. 1743–1752.
- [129] Li Y, Lin Z, Zhang S, Fu Q, Chen B, Lou JG, et al. Making language models better reasoners with step-aware Verifier. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 5315–5333. Available from: <https://aclanthology.org/2023.acl-long.291>.
- [130] Ding B, Qin C, Liu L, Bing L, Joty S, Li B. Is GPT-3 a good data annotator?; 2022. ArXiv:2212.10450.
- [131] Yoo KM, Park D, Kang J, Lee SW, Park W. GPT3Mix: leveraging large-scale language models for text augmentation. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021. p. 2225–2239.