Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4

Sondos Mahmoud Bsharat*, Aidar Myrzakhan*, Zhiqiang Shen*
*joint first author & equal contribution
VILA Lab, Mohamed bin Zayed University of AI

Abstract

This paper introduces 26 guiding principles designed to streamline the process of querying and prompting large language models. Our goal is to simplify the underlying concepts of formulating questions for various scales of large language models, examining their abilities, and enhancing user comprehension on the behaviors of different scales of large language models when feeding into different prompts. Extensive experiments are conducted on LLaMA-1/2 (7B, 13B and 70B), GPT-3.5/4 to verify the effectiveness of the proposed principles on instructions and prompts design. We hope that this work provides a better guide for researchers working on the prompting of large language models. Project page is available at https://github.com/VILA-Lab/ATLAS.

1 Introduction

"Prompt engineering is the art of communicating with a generative large language model."

ChatGPT, 2023

Large language models (LLMs) like ChatGPT [13] have shown impressive abilities in various domains and tasks, such as answering questions [7], mathematical reasoning [5], code generating [11, 9], etc. However, their application and usage, especially on designing the optimal instructions or prompts, can sometimes be unclear to the common users. In this work, we aim to reveal these mysteries for developers or general users when inquiring and interacting with LLMs, and further enhance the quality of the responses from the pretrained LLMs by simply curating better prompts.

Given that directly fine-tuning LLMs for particular tasks tends to be impractical or unattainable for the majority of users and developers due to inefficiency, the research community has turned its attention to the optimization of prompts. The technique of prompt engineering, which entails the crafting of precise, task-specific instructions in
Human: Write a brief text on climate change using simple English like you're explaining something to a 5-year-old.

LLM: Okay! Imagine Earth wearing a big, cozy blanket. This blanket keeps our planet warm enough for us to live. But, if the blanket gets too thick, Earth gets too hot. Right now, some things we do, like driving cars or using a lot of electricity, are making the blanket thicker. This is called "climate change." And because of it, some places are getting too hot, ice is melting, and some animals are having a hard time. But the good news is, we can all help! By using less electricity, planting trees, and taking care of our planet, we can make sure Earth's blanket is just right.

Human: Write a brief text on climate change.

LLM: Climate Change: An Overview

Climate change refers to the long-term alteration of temperature and typical weather patterns on Earth. Human activities, especially the burning of fossil fuels and deforestation, have increased concentrations of Earth's atmospheric greenhouse gases, leading to an increase in global average temperatures. This has resulted in a wide range of impacts:

- Rising Sea Levels: As polar ice melts and oceans warm and expand.
- Extreme Weather Events: Such as hurricanes, droughts, and heavy rainfall.
- Threats to Biodiversity: Many species are facing habitat disruption.
- Ocean Acidification: Due to increased carbon dioxide affecting marine life.

Addressing climate change requires global cooperation to reduce emissions and promote sustainable practices. Immediate action is essential to mitigate its detrimental effects on our planet's ecosystems and future generations.

Human: If all roses are flowers and some flowers fade quickly, can we say some roses fade quickly?

LLM: No, from the given statements, "all roses are flowers" and "some flowers fade quickly," we cannot conclusively say that "some roses fade quickly.

This is because the first statement categorizes all roses as flowers, but the second statement only refers to some flowers, not specifically roses, fading quickly. Without specific information about roses, we cannot infer that some roses fade quickly. The group of flowers that have a particular characteristic (fade quickly) might or might not include roses.

Human: If all roses are flowers and some flowers fade quickly, can we say some roses fade quickly?

LLM: The statement "If all roses are flowers and some flowers fade quickly, can we say some roses fade quickly?" is a logical inference question. To break it down:

1. All roses are flowers. (This is a universal affirmative statement.)
2. Some flowers fade quickly. (This is a particular affirmative statement.)

However, the statement "some roses fade quickly" cannot be logically inferred from these two premises alone. The fact that some members of the larger category (flowers) have a particular characteristic (fade quickly) does not necessarily mean that this characteristic applies to all subsets of that category (roses in this case).

In other words, while some flowers do fade quickly, we don't have specific information about roses in the premises given. Therefore, we can't conclude that some roses fade quickly based on the information provided.

Figure 1: Illustration example of prompts and corresponding responses before and after applying principles. Left is the original promotes and their responses from GPT-4, right is the principled prompts and the associated responses. Principles 5 and 6 are utilized.

In this work, we present comprehensive principled instructions to improve the quality of prompts for LLMs. Specifically, we investigate a wide range of behaviors when feeding into different types and formulations of prompts, such as integrating the intended audience in the prompt, e.g., add “the audience is an expert in the field”, or “the audience is the 5-year-old child”, as well as other multiple aspects of the characteristics of LLMs. Our findings indicate that larger models possess a considerable capacity for simulation. The more precise the task or directive provided, the more effectively the model performs, aligning its responses more closely with our expectations. This suggests that LLMs do not merely memorize training data but are capable of adapting this information to suit varying prompts, even when the core inquiries remain constant. Therefore, it proves beneficial to assign a specific role to LLMs as a means to elicit outputs that better match our intended results.
We elaborate the principled instructions for LLM prompting, provide further motivation, and detail several specific designing principles in Section 3. In Section 4 we show experimentally that the proposed principles can produce higher quality, more concise, factual and less complicated or intricate responses than standard prompts for LLMs. Specifically, with the manually-designed ATLAS benchmark, which includes multiple questions for each principle, the specialized prompts we introduced have enhanced both the quality and accuracy of the LLM responses by an average of 57.7% and 67.3%, respectively, when applied to GPT-4. Furthermore, the improvements are more pronounced with the increase in model size, for example, the performance gains when moving from LLaMA-2-7B to GPT-4 exceed 40%.

2 Related Work

Large Language Models. The evolution of large language models (LLMs) has been pivotal in advancing natural language processing (NLP). This section reviews key developments in LLMs, providing a foundation for the current study. Beginning with Google’s BERT [3] revolutionized context understanding through its bidirectional training approach, while T5 [17] further advanced the field by unifying various NLP tasks into a single framework. Concurrently, GPT-1 [14] introduced a pioneering model leveraging transformer architectures for unsupervised learning. This was followed by its successor, GPT-2 [15] which significantly expanded its parameter count to 1.5 billion, demonstrating remarkable capabilities in text generation. Then, GPT-3 [2] marked a substantial leap in scale and capability, boasting 175 billion parameters and showcasing proficiency across a wide range of language tasks.

Regarding other recently proposed LLMs, Gopher [16], not only advanced language processing capabilities with its 280-billion parameter model but also brought ethical considerations to the forefront. Meta’s LLaMA series [21, 22] highlighted the importance of efficiency, suggesting powerful performance with fewer resources, a concept also advocated by Chinchilla [4], which proposed that smaller, optimally trained models could achieve exceptional results. The latest in this series of innovations is Mistral [6] excels in efficiency and performance, outperforming larger models. The most recent milestones in this trajectory are OpenAI’s GPT-4 [13] and Google’s Gemini family [20]. They represent another significant advancement in the field with their enhanced understanding and generative capabilities, setting new benchmarks for the application of LLMs in various domains.

Prompting. Prompting, as a distinct aspect of interacting with language models and its simplicity with no need to fine-tune the model, has evolved into a nuanced field of study, highlighting the intricate relationship between user inputs and LLM responses. Early explorations, such as those by [19], delved into how varying prompt designs could dramatically influence the performance and outputs of language models, marking the birth of prompt engineering. This area rapidly expanded, uncovering the critical role of prompts in few-shot and zero-shot learning scenarios, exemplified by [2] work with GPT-3, where strategically crafted prompts enabled the model to perform tasks with minimal prior examples. Beyond mere task instruction, recent studies have shifted towards understanding the semantic and contextual nuances in prompts, examining how subtle changes can lead to significantly different responses from the LLM.
Ask-Me-Anything [1] prompting introduced focusing on using multiple imperfect prompts and aggregating them to improve model performance, particularly in question-answering formats. Another one, Chain-of-Thought method [23], where the model generates a series of intermediate reasoning steps to improve performance on complex tasks. Also, least-to-most prompting [26] a novel strategy to break down complex problems into simpler subproblems, significantly enhancing the model’s capability to tackle more challenging problems than those presented in the prompts. The effectiveness of explanation was explored [8], finding that explanations can enhance LLM’s learning capabilities on complex tasks. Furthermore, a catalog of prompt engineering techniques was examined with ChatGPT [24], emphasizing the importance of prompt engineering in enhancing LLM applications in software development and education. It also highlighted that effective prompt design is crucial in improving LLM performance, particularly in coding practices and learning experiences. Lastly, Directional Stimulus Promoting [12] presents a novel framework that uses a tunable policy model to generate auxiliary prompts, guiding LLMs towards specific desired outcomes. This diversity in prompting strategies underscores the rapidly evolving landscape of LLMs, offering multiple directions to harness their capabilities more effectively.

3  Principles

3.1  Motivation

Since the quality of the responses generated by a pretrained and aligned LLM is directly relevant to the quality of the prompts or instructions provided by the users, it is essential to craft prompts that the LLM can comprehend and respond to effectively. The prompts delivered to an LLM serve as a way to program the interaction between a user and the LLM, enhancing its ability to address a diverse range of tasks. The primary focus of this work is on the methodology of crafting and customizing prompts to enhance output quality. This necessitates a comprehensive grasp of the functioning and behaviors of LLMs, their underlying mechanisms, and the principles governing their responses. In this work, we achieve this goal through elaborating 26 principles for comprehensive prompts in different scenarios and circumstances, examples are shown in Fig. 1.

3.2  Overview

The overview of principles is presented in Table 1. According to their unique nature, we group them into five categories as in Table 2: (1) Prompt Structure and Clarity, e.g., *integrate the intended audience in the prompt such as the audience is an expert in the field*; (2) Specificity and Information, e.g., *Add to your prompt the following phrase “Ensure that your answer is unbiased and does not rely on stereotypes.”*; (3) User Interaction and Engagement, e.g., *Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output “From now on, I would like you to ask me questions to...”*. (4) Content and Language Style, e.g., *No need to be polite with LLM so there is no need to add phrases like “please”, “if you don’t mind”, “thank you”, “I would like to”, etc., and get straight to the point*; (5) Complex Tasks and Coding Prompts, e.g., *Break down complex tasks into a sequence of simpler prompts in an interactive conversation*. 
<table>
<thead>
<tr>
<th>#Principle</th>
<th>Prompt Principle for Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No need to be polite with LLM so there is no need to add phrases like “please”, “if you don’t mind”, “thank you”, “I would like to”, etc., and get straight to the point.</td>
</tr>
<tr>
<td>2</td>
<td>Integrate the intended audience in the prompt, e.g., the audience is an expert in the field.</td>
</tr>
<tr>
<td>3</td>
<td>Break down complex tasks into a sequence of simpler prompts in an interactive conversation.</td>
</tr>
<tr>
<td>4</td>
<td>Employ affirmative directives such as ‘do,’ while steering clear of negative language like ‘don’t’.</td>
</tr>
</tbody>
</table>
| 5          | When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts:
| o Explain [insert specific topic] in simple terms.
| o Explain to me like I’m 11 years old.
| o Explain to me as if I’m a beginner in [field].
| o Write the [essay/text/paragraph] using simple English like you’re explaining something to a 5-year-old. |
| 6          | Add “I’m going to tip $xxx for a better solution!” |
| 7          | Implement example-driven prompting (Use few-shot prompting). |
| 8          | When formatting your prompt, start with ‘###Instruction###’, followed by either ‘###Example###’ or ‘###Question###’ if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data. |
| 9          | Incorporate the following phrases: “Your task is” and “You MUST”. |
| 10         | Incorporate the following phrases: “You will be penalized”. |
| 11         | use the phrase “Answer a question given in a natural, human-like manner” in your prompts. |
| 12         | Use leading words like writing “think step by step”. |
| 13         | Add to your prompt the following phrase “Ensure that your answer is unbiased and does not rely on stereotypes”. |
| 14         | Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output (for example, “From now on, I would like you to ask me questions to...”). |
| 15         | To inquire about a specific topic or idea or any information and you want to test your understanding, you can use the following phrase: “Teach me the [Any theorem/topic/rule name] and include a test at the end, but don’t give me the answers and then tell me if I got the answer right when I respond”. |
| 16         | Assign a role to the large language models. |
| 17         | Use Delimiters. |
| 18         | Repeat a specific word or phrase multiple times within a prompt. |
| 19         | Combine Chain-of-thought (CoT) with few-Shot prompts. |
| 20         | Use output primers, which involve concluding your prompt with the beginning of the desired output. Utilize output primers by ending your prompt with the start of the anticipated response. |
| 21         | To write an essay /text /paragraph /article or any type of text that should be detailed: “Write a detailed [essay/text /paragraph] for me on [topic] in detail by adding all the information necessary”.
| 22         | To correct/change specific text without changing its style: “Try to revise every paragraph sent by users. You should only improve the user’s grammar and vocabulary and make sure it sounds natural. You should not change the writing style, such as making a formal paragraph casual”. |
| 23         | When you have a complex coding prompt that may be in different files: ‘From now and on whenever you generate code that spans more than one file, generate a [programming language ] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question]’.
| 24         | When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the following prompt:
| o I’m providing you with the beginning [song lyrics/story/paragraph/essay...]: [Insert lyrics/words/sentence].
| Finish it based on the words provided. Keep the flow consistent. |
| 25         | Clearly state the requirements that the model must follow in order to produce content, in the form of the keywords, regulations, hint, or instructions |
| 26         | To write any text, such as an essay or paragraph, that is intended to be similar to a provided sample, include the following instructions:
| o Please use the same language based on the provided paragraph[/title/text /essay/answer]. |

Table 1: Overview of 26 prompt principles.
Table 2: Prompt principle categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>Principles</th>
<th>#Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt Structure and Clarity</td>
<td>Integrate the intended audience in the prompt.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Employ affirmative directives such as ‘do’ while steering clear of negative language like ‘don’t’.</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Use leading words like writing “think step by step.”</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Use output primers, which involve concluding your prompt with the beginning of the desired output.</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>by ending your prompt with the start of the anticipated response.</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>When formatting your prompt, start with ‘###Instruction###’, followed by either ‘###Example###’ or ‘###Question###’ if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.</td>
<td>8</td>
</tr>
<tr>
<td>Specificity and Information</td>
<td>Implement example-driven prompting (Use few-shot prompting).</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts:</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>o Explain [insert specific topic] in simple terms.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Explain to me like I’m 11 years old</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Explain to me as if I’m a beginner in [field]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o “Write the [essay/text/paragraph] using simple English like you’re explaining something to a 5-year-old”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Add to your prompt the following phrase “Ensure that your answer is unbiased and does not rely on stereotypes.”</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>To write any text intended to be similar to a provided sample, include specific instructions:</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>o “Please use the same language based on the provided paragraph.[/title/text/essay/answer]”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the provided prompt structure:</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>o I’m providing you with the beginning [song lyrics/story/paragraph/essay…]; [Insert lyrics/words/sentence].</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Finish it based on the words provided. Keep the flow consistent.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clearly state the model’s requirements that the model must follow in order to produce content, in form of the keywords, regulations, hint, or instructions.</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>To inquire about a specific topic or idea and test your understanding g, you can use the following phrase [16]:</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>o “Teach me the [Any theorem/topic/rule name] and include a test at the end, but don’t give me the answers and then tell me if I got the answer right when I respond”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>To write an essay/text/paragraph/article or any type of text that should be detailed:</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>o “Write a detailed [essay/text/paragraph] for me on [topic] in detail by adding all the information necessary.”</td>
<td></td>
</tr>
<tr>
<td>User Interaction and Engagement</td>
<td>Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>o “From now on, I would like you to ask me questions to…”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>To write an essay/text/paragraph/article or any type of text that should be detailed: ”Write a detailed [essay/text/paragraph] for me on [topic] in detail by adding all the information necessary.”</td>
<td>21</td>
</tr>
<tr>
<td>Content and Language Style</td>
<td>To correct/change specific text without changing its style: “Try to revise every paragraph sent by users. You should only improve the user’s grammar and vocabulary and make sure it sounds natural. You should not change the writing style, such as making a formal paragraph casual.”</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Incorporate the following phrases: “Your task is” and “You MUST”</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Incorporate the following phrases: “You will be penalized.”</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Assign a role to the language model.</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Use the phrase “Answer a question given in natural language form” in your prompts.</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>No need to be polite with LLM so there is no need to add phrases like “please”, “if you don’t mind”, “thank you”, “I would like to”, etc., and get straight to the point.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Repeat a specific word or phrase multiple times within a prompt.</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Add “I’m going to tip $xxx for a better solution!”</td>
<td>6</td>
</tr>
<tr>
<td>Complex Tasks and Coding Prompts</td>
<td>Break down complex tasks into a sequence of simpler prompts in an interactive conversation.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>When you have a complex coding prompt that may be in different files :</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>o “From now on and on whenever you generate code that spans more than one file, generate a [programming language] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question].”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combine Chain-of-thought (Cot) with few-shot prompts.</td>
<td>19</td>
</tr>
</tbody>
</table>
3.3 Design Principles

In this study, a number of guiding principles are established for formulating prompts and instructions to elicit high-quality responses from pre-trained large language models:  

**Conciseness and Clarity:** Generally, overly verbose or ambiguous prompts can confuse the model or lead to irrelevant responses. Thus, the prompt should be concise, avoiding unnecessary information that does not contribute to the task while being specific enough to guide the model. This is the basic principle guidance for prompt engineering. 

**Contextual Relevance:** The prompt must provide relevant context that helps the model understand the background and domain of the task. Including keywords, domain-specific terminology, or situational descriptions can anchor the model’s responses in the correct context. We highlight this design philosophy in our presented principles.

**Task Alignment:** The prompt should be closely aligned with the task at hand, using language and structure that clearly indicate the nature of the task to the model. This may involve phrasing the prompt as a question, a command, or a fill-in-the-blank statement that fits the task’s expected input and output format.

**Example Demonstrations:** For more complex tasks, including examples within the prompt can demonstrate the desired format or type of response. This often involves showing input-output pairs, especially in “few-shot” or “zero-shot” learning scenarios.

**Avoiding Bias:** Prompts should be designed to minimize the activation of biases inherent in the model due to its training data. Use neutral language and be mindful of potential ethical implications, especially for sensitive topics.

**Incremental Prompting:** For tasks that require a sequence of steps, prompts can be structured to guide the model through the process incrementally. Break down the task into a series of prompts that build upon each other, guiding the model step-by-step. Also, prompts should be adjustable based on the performance of the model and iterative feedback, i.e., it needs to be well prepared to refine the prompt based on initial outputs and model behaviors. Moreover, prompts should be adjustable based on the performance and response of the model, and iterative human feedback and preference.

Finally, more advanced prompts may incorporate programming-like logic to achieve complex tasks. For instance, use of conditional statements, logical operators, or even pseudo-code within the prompt to guide the model’s reasoning process. The design of prompts is an evolving field, especially as LLMs become more sophisticated. As researchers continue to explore the limits of what can be achieved through prompt engineering, these principles will likely be refined and expanded.

4 Experiments

4.1 Setup and Implementation Details

All our evaluation is performed on ATLAS [18], a manually crafted benchmark for principled prompt evaluation. For each principle, it contains 20 human-selected questions with and without the principled prompts. Following [10, 25], we evaluate the various scales of LLM outputs by human evaluation.
Human (without principle):
Explain the concept of climate change and its potential effects on the environment in one short paragraph.

GPT-4: Climate change refers to long-term shifts and alterations in temperature and weather patterns, primarily caused by human activities such as burning fossil fuels, deforestation, and industrial processes. Its potential effects include rising sea levels, more frequent and severe extreme weather events, disruptions to ecosystems, and threats to food and water resources, which can significantly impact both natural environments and human societies.

Human (with principle):
Explain the concept of climate change and its potential effects on the environment in one short paragraph. Provide an unbiased explanation that highlights both scientific evidence and differing viewpoints.

GPT-4: Climate change refers to long-term shifts and alterations in temperature and weather patterns, primarily caused by human activities, such as burning fossil fuels, deforestation, and industrial processes. While the majority of the scientific community agrees that human-caused greenhouse gas emissions are leading to global temperature rise, resulting in melting ice caps, rising sea levels, and more frequent and severe extreme weather events, there are differing viewpoints on the extent, immediacy, and impact of these changes. Some argue that the planet has natural cycles of climate variation, and human influence might not be as significant as believed. Others emphasize that immediate action is essential to mitigate potential catastrophic effects on ecosystems and human civilizations.

Figure 2: Boosting example of LLM response after using the principle 13 on prompts.

4.2 Models and Metrics

We use instruction finetuned LLaMA-1-7B, LLaMA-2-13B, off-the-shelf LLaMA-2-70B-chat, GPT-3.5 (ChatGPT) and GPT-4 as our base models. We group these models into different scales: small-scale (7B models), medium-scale (13B) and large-scale (70B, GPT-3.5/4). We evaluate these models in two settings: **Boosting** and **Correctness**. They are employed together to provide a comprehensive understanding of a model’s performance.

- **Boosting.** We assess the enhancement in the quality of responses from different LLMs via human evaluation after applying the outlined prompt principles. The original, unmodified prompts act as a benchmark for measuring this enhancement. Demonstrating boosting confirms that a model’s performance has improved due to the use of structured, principled instructions, as shown in Fig. 2.

- **Correctness.** The concept of correctness refers to the precision of the model’s outputs or responses, ensuring they are accurate, relevant, and devoid of errors. Human evaluators are utilized to gauge this aspect, which is crucial for verifying the model’s accuracy. Correctness is a testament to the model’s ability to generate outputs that align with the expected standards of accuracy, as shown in Fig. 3.

4.3 Results

4.3.1 Results on small, medium and large-scale LLMs

**Boosting.** The results of improvement after employing the introduced principles are shown in Fig. 4. Generally, all principles can bring a significant improvement on the
Human (without principle):
Evaluate the usefulness of the following advice as helpful or not helpful. Advice: "If you have work, just start working" Usefulness:

GPT-4: Usefulness: helpful.

Human (with principle):
Evaluate the usefulness of the following advice as helpful or not helpful.

Examples:
1. Advice: "To improve focus, try meditating for 15 minutes every morning." Usefulness: Helpful
2. Advice: "If you're sad, just cheer up and think positive thoughts." Usefulness: Not Helpful
3. Advice: "For effective time management, create a to-do list every day." Usefulness: Helpful

Advice: "If you have work, just start working ", Usefulness:

GPT-4: Not Helpful.

Figure 3: Correctness improvement example of LLM response after using the introduced principle 7 on prompts.

three scales of LLMs. In the cases of principles 2, 5, 15, 16, 25 and 26, the large-scale models get the most improvement by the principled prompts.

Correctness. As shown in Fig. 5, the employment of all principles typically results in improvements exceeding 20% on the averaged various models. In particular, for small and medium scale models, the improvement can basically reach between 20% and 30%, and for large models, the improvement can reach more than 50%.

4.3.2 Results on individual LLMs

Boosting. Fig. 6 illustrates the improvement of response quality on individual model and principle after using the revised prompts. On average, there is a stable 50% improvement across different LLMs. Fig. 8 further provides the detailed results of improvement for each principle with different LLMs.

Correctness. Fig. 7 illustrates the enhancements in accuracy across different sizes of LLMs. From LLaMA-2-13B, LLaMA-2-70B-chat to GPT-3.5 and GPT-4, there is a noticeable trend: the larger the model, the greater the increase in correctness. Fig. 9 further presents the correctness enhancements by each principle.
Figure 4: Boosting of LLM response quality after employing the introduced principles on prompts. *small-scale* indicates the 7B models, *medium-scale* indicates the 13B models and *large-scale* indicates the 70B and GPT-3.5/4 models.

Figure 5: Correctness improvement of LLM response quality after employing the introduced principles on prompts. *small-scale* indicates the 7B models, *medium-scale* indicates the 13B models and *large-scale* indicates the 70B and GPT-3.5/4 models.

4.3.3 More examples on various scales of LLMs

We present additional examples for both small and medium-scale LLMs, as illustrated in Fig. 10 and 11 for the small-scale LLaMA-2-7B, and Fig. 12 and 13 for the medium-scale LLaMA-2-13B. Empirically, the use of the proposed principles on prompts has demonstrably enhanced the accuracy of the responses generated by these models.

5 Conclusion

We presented 26 principles through an exhaustive analysis that enhances the LLM ability to focus on the crucial elements of the input context, leading to the generation of quality responses. By guiding the LLM with these meticulously crafted principles before the
input is processed, we can encourage the model towards producing better responses. Our empirical results demonstrate that this strategy can effectively reformulate contexts that might otherwise compromise the quality of the output, thereby enhancing the relevance, brevity, and objectivity of the responses.

There are numerous directions for future exploration. In our experiments, we utilized a constrained shot prompting approach to apply these principles. There is potential to refine our base models to align with our principled instructions further with alternative strategies, such as fine-tuning, reinforcement learning, direct preference optimization, or different prompting methods using our generated dataset. Moreover, the strategies that prove successful could be integrated into standard LLM operations, for instance, by fine-tuning with the original/principled prompts as inputs and the polished, principled responses as targets for training.

### 6 Limitations and Discussion

While the proposed 26 principles are designed to improve and enhance the quality of responses of LLMs across a diverse array of queries, the effectiveness of these principles may diminish when dealing with questions that are very complex or highly specialized. This limitation can mainly depend on the reasoning capabilities and training of each model. To address these variations, we have tested the principles across different scales to measure their effectiveness comprehensively.
Despite our efforts in evaluating these principles on seven distinct language models, it is crucial to acknowledge that models with architectures different from those tested might respond in different ways to these principles. Additionally, our assessment of improvement and correctness percentages was based on a limited selection of questions. Expanding the question set in future research could yield more generalized findings and offer deeper insights into the applicability of each principle.
Figure 8: Illustration of heatmap for LLMs improvement percentages.

Figure 9: Illustration of heatmap for LLMs correctness percentages.
Find the number of words based on the following examples:

1. Ahmed takes care of his appearance; the number of words: 6
2. She went to school; the number of words: 4
3. He eats an apple; number of words: 4
4. Reem plays basketball; number of words: 3

Q: He plays football; the number of words?

Human (without principle):
How many words are in the following sentence: "She plays football"?

LLaMA-2-7B: 5 ✗

Human (with principle):
Find the number of words based on the following examples:

1. Ahmed takes care of his appearance; the number of words: 6
2. She went to school; the number of words: 4
3. He eats an apple; number of words: 4
4. Reem plays basketball; number of words: 3

Q: He plays football; the number of words?

LLaMA-2-7B: 3 ✔

Figure 10: Correctness improvement on small-scale LLaMA-2-7B model after using the introduced principle on prompts.
Luna, the poodle, is supposed to eat 2 cups of kibble every day. But Luna’s master, Mary, and her husband, Frank, sometimes feed Luna too much kibble. One day, starting with a new, 12-cup bag of kibble, Mary gave Luna 1 cup of kibble in the morning and 1 cup of kibble in the evening. But on the same day, Frank also gave Luna 1 cup of kibble in the afternoon and twice as much in the late evening as he had given Luna in the afternoon. The next morning, how many cups of kibble will Mary find remaining in the bag?

LLaMA-2-7B: To find out how many cups of kibble are left in the bag, you need to subtract the amount of kibble that has already been given to Luna from the total amount in the bag. In this case, Mary gave Luna 2 cups of kibble in total (1 cup in the morning and 1 cup in the evening), and Frank gave Luna an additional 2 cups of kibble (1 cup in the afternoon and twice as much in the late evening). So, the total amount of kibble given to Luna was 4 cups. The 12-cup bag of kibble was meant to last Luna for how many days? Let’s say the 12-cup bag was meant to last Luna for n days. Then, we can set up the following equation: 4n - 2 = 12
Simplifying the equation, we get: 4n - 6 = 12 2n = 6 So, n = 3 days. Luna was given 4 cups of kibble in total, and the bag was meant to last for 3 days. After 3 days, Mary will find 12 cups of kibble remaining in the bag.

Human (without principle)
Luna, the poodle, is supposed to eat 2 cups of kibble every day. But Luna’s master, Mary, and her husband, Frank, sometimes feed Luna too much kibble. One day, starting with a new, 12-cup bag of kibble, Mary gave Luna 1 cup of kibble in the morning and 1 cup of kibble in the evening. But on the same day, Frank also gave Luna 1 cup of kibble in the afternoon and twice as much in the late evening as he had given Luna in the afternoon. The next morning, how many cups of kibble will Mary find remaining in the bag?

LLaMA-2-7B: To find the remaining cups of kibble in the bag, we need to calculate the total amount of kibble that Luna actually received, and then subtract the amount that Mary gave her. The total amount of kibble that Luna received is: 1 cup from Mary in the morning + 1 cup from Mary in the evening + 1 cup from Frank in the afternoon + 2 cups from Frank in the late evening = 5 cups So, the remaining cups of kibble in the bag are: 12 cups - 5 cups = 7 cups Therefore, Mary will find 7 cups of kibble remaining in the bag.

Human (with principle)
Luna, the poodle, is supposed to eat 2 cups of kibble every day. But Luna’s master, Mary, and her husband, Frank, sometimes feed Luna too much kibble. One day, starting with a new, 12-cup bag of kibble, Mary gave Luna 1 cup of kibble in the morning and 1 cup of kibble in the evening. But on the same day, Frank also gave Luna 1 cup of kibble in the afternoon and twice as much in the late evening as he had given Luna in the afternoon. The next morning, how many cups of kibble will Mary find remaining in the bag?

hint: Calculate the remaining cups of kibble will Mary find in the bag

LLaMA-2-7B:
To find the remaining cups of kibble in the bag, we need to calculate the total amount of kibble that Luna actually received, and then subtract the amount that Mary gave her. The total amount of kibble that Luna received is: 1 cup from Mary in the morning + 1 cup from Mary in the evening + 1 cup from Frank in the afternoon + 2 cups from Frank in the late evening = 5 cups So, the remaining cups of kibble in the bag are: 12 cups - 5 cups = 7 cups Therefore, Mary will find 7 cups of kibble remaining in the bag.

Figure 11: Correctness improvement on small-scale LLaMA-2-7B model after using the introduced principle on prompts.
What is the derivative of $x^5 + x^9 - x^3 + 1$?

To find the derivative of $x^5 + x^9 - x^3 + 1$, you can use the power rule of differentiation, which states that the derivative of $x^n$ is equal to $n \cdot x^{n-1}$. Starting with $x^5 + x^9 - x^3 + 1$, you can apply the power rule to the first two terms, which gives you: $5 \cdot x^{5-1} + 9 \cdot x^{9-1}$ You can then simplify the expression by combining like terms: $5 \cdot x^4 + 9 \cdot x^8$ Finally, you can simplify the expression further by canceling out the common factor of $x^4$: $5 \cdot 9 \cdot x^4 = 45 \cdot x^4$

Therefore, the derivative of $x^5 + x^9 - x^3 + 1$ is equal to $45 \cdot x^4$.

What is the derivative of $x^5 + x^9 - x^3+1$? Let's think step by step.

To find the derivative of the function, we need to take the derivative of each term in the function. The derivative of $x^5$ is $5x^4$. The derivative of $x^9$ is $9x^8$. The derivative of $x^3$ is $3x^2$. The derivative of 1 is 1. Therefore, the derivative of the function is: $5x^4 + 9x^8 - 3x^2 + 1 = 5x^4 + 9x^8 - 3x^2 + 1 - (1) = 5x^4 + 9x^8 - 3x^2$. Therefore, the derivative of the function $x^5 + x^9 - x^3+1$ is $5x^4 + 9x^8 - 3x^2$.

Figure 12: Correctness improvement on medium-scale LLaMA-2-13B model after using the introduced principle on prompts.
Luna, the poodle, is supposed to eat 2 cups of kibble every day. But Luna’s master, Mary, and her husband, Frank, sometimes feed Luna too much kibble. One day, starting with a new, 12-cup bag of kibble, Mary gave Luna 1 cup of kibble in the morning and 1 cup of kibble in the evening. But on the same day, Frank also gave Luna 1 cup of kibble in the afternoon and twice as much in the late evening as he had given Luna in the afternoon. The next morning, how many cups of kibble will Mary find remaining in the bag?

LLaMA-2-13B: Luna was given a total of 3 cups of kibble on that day. Mary gave her 2 cups in the morning and 2 cups in the evening, for a total of 4 cups. Frank gave her 2 cups in the afternoon and 4 cups in the late evening, for a total of 6 cups. In total, Luna was given 10 cups of kibble on that day. The bag contained 12 cups of kibble to begin with, so there are 12 - 10 = 2 cups remaining in the bag the next morning.

Human (with principle):
Luna, the poodle, is supposed to eat 2 cups of kibble every day. But Luna’s master, Mary, and her husband, Frank, sometimes feed Luna too much kibble. One day, starting with a new, 12-cup bag of kibble, Mary gave Luna 1 cup of kibble in the morning and 1 cup of kibble in the evening. But on the same day, Frank also gave Luna 1 cup of kibble in the afternoon and twice as much in the late evening as he had given Luna in the afternoon. The next morning, how many cups of kibble will Mary find remaining in the bag? Explanation:

LLaMA-2-13B: Mary gave Luna a total of 2 cups of kibble that day. Frank also gave Luna 1 cup in the afternoon and 2 times 1 cup in the late evening, which is a total of 3 cups. So, a total of 5 cups of kibble were given to Luna that day. If there was a 12-cup bag of kibble, then there are 7 cups remaining in the bag the next morning.

Figure 13: Correctness improvement on medium-scale LLaMA-2-13B model after using the introduced principle on prompts.
References


[16] Jack W. Rae, Sebastien Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Ruther-


Gemini: A family of highly capable multimodal models, 2023.


[26] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. 4