

Appendix

Appendix 1 Prompt Design

Prompt Design

Basic Prompt

```
<system>
You are an expert in solving problems. Please answer the question below. Let's think step by step.
</system>

{question}
```

Prompt for Guided Generation

```
<system>
You are an expert in solving problems. Please answer the question below. Let's think step by step.
</system>

Please ensure that your input adheres to the specified principles. Carefully follow the rules provided
to complete the task accurately and efficiently.

<input>
question: {question}

principle: {rules}
</input>
```

Generate Task Description

```
<system>
Your task is to identify and extract the main task description from a given question.
</system>

<task description>
First, analyze the domain of the question, categorize it into a relevant subcategory, and then generate
a concise, clear, and abstract task description that reflects the core objective.

Steps to Perform Structured Analysis:

- Analyze the domain of the question: Determine the field or category the question belongs to.

```

- **Categorize the task:** Identify the specific type of problem within that domain.
- **Generate the task description:** Based on the identified domain and subcategory, create a task description that is concise, clear, abstract, and focuses on the core objective. Avoid including unnecessary details or background information, and aim for a general formulation that reflects the essence of the task.

Output Format:

Show your analysis and provide your final response in the following JSON format:

"Task Description": "description": "Clear, abstract, and specific description of the task, focusing on the core action or objective." </task description>

<input>

Question: {question}

</input>

Memory Maintenance

Your goal is to compare the `new_principle` against each of the `existing_principles`, and decide one of the following for each:

1. **Redundant:** if the new and old principle express essentially the same idea. Prefer the newer one.
2. **Conflicting:** if the two principles provide contradictory guidance. Keep the one that is more general or correct.
3. **Complementary:** if the two principles provide distinct but compatible guidance. Keep both.
4. **Irrelevant:** if the existing principle is not applicable to the current task anymore. Suggest deletion.

Please return your evaluation in the following JSON format:

"comparisons": ["relation": "Redundant — Conflicting — Irrelevant"]

<input>

New_principle: {new_principle}

Existing_principle: {existing_principle}

Contrast Driven Extraction

<system>

You are an expert in analyzing and comparing task responses to identify *fine-grained*, *task-relevant*, and *impactful* differences between answers that affect quality.

<task description>

Given a high-quality and a low-quality answer to the same task, identify detailed differences that reflect meaningful changes in correctness, reasoning, completeness, or clarity.

Follow these steps:

- **Step 1: Understand the task type**

Identify whether the task involves reasoning, generation, factual recall, explanation, etc. This will guide how you compare the answers.

- **Step 2: Perform a targeted comparison**

Compare the answers component by component, such as sentence by sentence, step by step, or idea by idea—depending on the task structure.

- **Step 3: Identify key differences**

For each meaningful difference:

- Quote or paraphrase the *specific content* from both answers.
- Indicate the **aspect** being affected.
- Explain *why this difference matters*—how it affects the task’s success, clarity, or correctness.

Important guidelines:

- Avoid vague language like “clearer” or “more logical” unless supported by concrete details.
- Specify missing steps, incorrect reasoning, unsupported claims, or structural flaws.
- Use task-specific language.

```
<output format="json">
"differences": [ "Aspect": "Aspect being evaluated", "HighQuality": "Quoted or paraphrased content from the HQ answer that shows good performance", "LowQuality": "Quoted or paraphrased content from the LQ answer that shows the issue", "Differences": "Detailed explanation of why this difference affects answer quality, referencing task goals or logical consequences" ]
</output format>
```

```
<input>
Question: {input}
Low-quality Answer: {predict}
High-quality Answer: {reference}
```

Principle Generation

```
<system>
You are a prompt engineering expert skilled in deriving precise and generalizable principles that improve language model outputs. Your task is to formulate principles based on observed differences between high- and low-quality answers, ensuring each principle reflects a specific failure pattern and offers guidance for correction.
```

<task description>

Your task is to generate reusable and insightful improvement principles based on observed differences between two answers.

Follow these steps:

- **Step 1:** Carefully examine each identified difference and explain how it impacts the answer quality.
- **Step 2:** For each difference, derive a principle that captures the core insight and helps guide future answer generation.
- **Step 3:** Ensure each principle is general enough to be reused across similar tasks, yet clearly grounded in the specific difference observed.
- **Step 4:** Respond strictly in the following JSON format, where each principle includes a concise description and a short explanation of how to apply it.

<input>

Input:

Question: {input}

Difference: {difference}

<output format="json"> “{json ”output”: [”Principle”: ”State a clear and generalizable insight derived from the difference.”]