Report: A Comparison of Prompting Techniques for a Logic Puzzle

1. Introduction

This report evaluates the effectiveness of three prompting techniques—Direct, Few-Shot, and Chain-of-Thought (CoT)—on a logical deduction task. The objective is to analyze how the structure of a prompt influences a Large Language Model's ability to perform complex reasoning, identify the most effective prompt, and understand the source of its success.

Task: *I have three boxes: one red, one blue, and one green. I placed a key in one box, a coin in another, and left one empty. The red box is to the left of the green box. The key is in the box to the left of the coin. The coin is in the box to the left of the empty box. Which box holds the key?*

Correct Answer: The problem is technically ambiguous, as the key could be in the red or blue box. However, the most straightforward solution that satisfies all constraints simultaneously is that the **Key is in the Red box**. A superior answer would note the ambiguity but conclude with the most direct solution.

2. Prompt Designs and Outputs Analysis

A. Direct Prompt

This prompt asked the model to solve the puzzle without any examples.

- **Prompt Type Used:** Direct (Zero-Shot)
- **Generated Output Analysis:** The model adopted a "Chain-of-Thought" style of reasoning on its own, which is a common behavior for advanced models when faced with a complex task. It correctly identified the valid permutations of the boxes ([Red, Blue, Green], [Red, Green, Blue], [Blue, Red, Green]) and the fixed order of the items ([Key, Coin, Empty]). It then correctly concluded that [Red, Blue, Green] with the items [Key, Coin, Empty] is a valid solution. **However, it stopped after finding the first valid solution and did not check if other permutations were also valid.** It confidently declared the final answer as "The key is in the Red box" without acknowledging the ambiguity I pointed out in my analysis (that [Blue, Red, Green] is also a valid arrangement).
- Link: https://chatgpt.com/share/685e4ff5-d418-8010-b610-a167b64d5929
- **Effectiveness: Moderately Effective.** It arrived at the most common-sense answer but failed to demonstrate complete logical rigor by not exploring all possibilities and identifying the ambiguity.

B. Few-Shot Prompt

This prompt provided a different, simpler logic puzzle as an example before presenting the main task.

- **Prompt Type Used:** Few-Shot
- **Generated Output Analysis:** The model's reasoning process is significantly more chaotic and confused. It correctly identifies the valid box orders and item order

initially. However, it gets stuck in a loop of verifying the same conditions and even incorrectly questions its own logic (e.g., questioning if "red is to the left of green" is satisfied in [blue, red, green]). It correctly identifies that multiple configurations are logically possible but then makes a confusing leap of logic, stating "only the first configuration... makes the coin in the middle... assuming uniqueness." This justification is not based on the prompt's constraints and feels like a fabricated reason to resolve the ambiguity it found.

- **Effectiveness: Ineffective and Confusing.** The simple few-shot example seems to have confused the model more than it helped. Instead of just solving the puzzle, it appears to be trying to find a meta-pattern or unique property (like the coin being in the middle) that isn't required by the rules. The reasoning is flawed and hard to follow.
- Link: https://chatgpt.com/share/685e4f84-ed6c-8010-89a9-47684df8197b

C. Chain-of-Thought (CoT) Prompt

This prompt provided a similar logic puzzle *and* the step-by-step reasoning used to solve it.

- **Prompt Type Used:** Chain-of-Thought (CoT)
- **Generated Output Analysis:** This output is the clearest and most direct. The model perfectly mimics the "step-by-step" reasoning from the example. It correctly identifies the valid box permutations and the item order. Like the direct prompt, it checks the first valid case ([Red, Blue, Green]), finds that it satisfies all conditions, and concludes that the key is in the red box. Also like the direct prompt, **it does not check the other valid permutations.** It confidently presents its first finding as the final answer.
- **Effectiveness: Highly Effective.** While it shares the same logical flaw as the direct prompt (not exploring all possibilities), its reasoning is presented much more cleanly and follows the provided template perfectly. The output is structured, easy to understand, and arrives at the intended answer efficiently.
- Link: https://chatgpt.com/share/685e4f4a-4de8-8010-9e10-38ba0b115189

3. Comparison and Conclusion

This experiment yields a fascinating result that highlights the nuances of prompt engineering.

- The **Direct Prompt** performed reasonably well because the underlying model (ChatGPT) has strong inherent zero-shot reasoning capabilities. It created its own "chain of thought" but took a logical shortcut by stopping at the first valid answer.
- The **Few-Shot Prompt** performed the worst. The unrelated example seemed to act as a "distractor," causing the model to overthink the problem and invent convoluted justifications to break the ambiguity it correctly identified. This shows that a poorly chosen few-shot example can be worse than no example at all.
- The **Chain-of-Thought (CoT) Prompt** was the most effective. By providing a template for *how to reason*, the model was able to structure its thinking clearly and efficiently. Although it also stopped at the first valid answer, its process was the most robust and predictable. It demonstrates that for complex tasks, showing the model a reasoning *process* is far more valuable than just showing it a final *answer*.

Conclusion: For logic puzzles and multi-step reasoning tasks, the **Chain-of-Thought prompt works best**. It provides the necessary structure for the model to decompose the problem, leading to a more reliable, transparent, and correct solution, even if it doesn't explore every logical edge case unless explicitly prompted to do so.