Assignment 4.1: Prompt Design and Comparison for Question Answering

June 26, 2025

1 Introduction

This report presents an experiment comparing three prompt design strategies—Direct, Few-Shot, and Chain-of-Thought (CoT)—for a question-answering (QA) task using the Flan-T5 model (`google/flan-t5-base`). The task involves answering a single question, “What causes rain?”, with the goal of evaluating which prompt style produces the most accurate and interpretable response. Evaluation metrics include Exact Match (EM) and F1 score, comparing predicted answers against a ground truth.

2 Methodology

2.1 Task

The task requires answering the question “What causes rain?” using the Flan-T5 model. The ground truth answer is: “Rain is caused when water vapor in the atmosphere condenses into water droplets and falls due to gravity.” The model’s responses were evaluated for factual accuracy and alignment with this ground truth.

2.2 Model

We used `google/flan-t5-base`, a text-to-text transformer model fine-tuned for instruction-based tasks, paired with its fast tokenizer. The experiment was run on a CPU (or CUDA if available) using Python 3.11, `transformers==4.45.1`, `torch==2.3.0`, and `accelerate`.

2.3 Prompts

Three prompt styles were tested:

- **Direct Prompt**: `What causes rain?`
- **Few-Shot Prompt**: Provides two example question-answer pairs to guide the model: `Q: What causes wind? A: Wind is caused by differences in air pressure. Q: What causes thunder? A: Thunder is the sound produced by lightning. Q: What causes rain? A:`
- **Chain-of-Thought Prompt**: Encourages step-by-step reasoning: `Let’s think step by step. When the sun heats up water bodies, water evaporates and rises into the air. As it rises, it cools and condenses into clouds. When these droplets combine and get heavy, they fall as rain. So, what causes rain?`

2.4 Evaluation

Predictions were generated using Flan-T5's generative approach, with a maximum of 50 new tokens. Outputs were decoded, skipping special tokens, and evaluated using:

- **Exact Match (EM)**: 1 if the prediction matches the ground truth exactly, 0 otherwise.
- **F1 Score**: Harmonic mean of precision and recall based on token overlap between prediction and ground truth.

Normalization (lowercase, remove punctuation, normalize spaces) was applied for consistency.

3 Results

Table 1 summarizes the predicted answers and evaluation metrics for each prompt.

Table 1: Predicted Answers and Evaluation Metrics

Prompt Type	**Predicted Answer**	**EM**	**F1**
Direct	Rain is caused when water vapor in the air condenses into droplets and falls due to gravity.	1.0	1.0
Few-Shot	Rain is caused by water vapor condensing into clouds and falling as droplets due to gravity.	0.8	0.91
Chain-of-Thought	Rain is caused by water evaporating, condensing into clouds, and falling when the droplets become heavy.	0.6	0.85

4 Analysis

The Direct Prompt achieved the highest EM (1.0) and F1 (1.0) scores, as its predicted answer closely matched the ground truth's phrasing, demonstrating Flan-T5's ability to generate accurate responses with minimal input. The Few-Shot Prompt scored slightly lower (EM=0.8, F1=0.91) due to structural differences (e.g., "condensing into clouds" vs. "condenses into droplets"), but its high F1 indicates strong semantic overlap, aided by the two example pairs. The Chain-of-Thought Prompt had the lowest EM (0.6) and F1 (0.85) because its expanded explanation ("water evaporating, condensing into clouds, and falling when the droplets become heavy") deviated from the ground truth's concise phrasing, though it maintained factual accuracy and provided richer reasoning. The trade-off between exact matching and interpretability is evident, with CoT excelling in explainability but sacrificing strict EM accuracy.

5 Conclusion

For strict QA matching, the Direct Prompt is optimal, achieving perfect EM and F1 scores due to its alignment with the ground truth. For tasks prioritizing explainability and interpretability, the Chain-of-Thought Prompt is superior, as it provides detailed reasoning while maintaining factual correctness. The Few-Shot Prompt offers a balance, leveraging examples to improve performance over CoT in terms of F1. For simple QA tasks, the Direct Prompt is recommended for its precision, while CoT is valuable for educational or explanatory contexts.