AI report

Two main performance metrics were used to assess various prompt engineering techniques: accuracy and response time. Accuracy is crucial for ensuring the reliability of this game, while response time significantly impacts user experience. In this analysis, we evaluate the model on a binary classification problem to determine whether the provided I'rab is correct or incorrect. The dataset was gathered from multiple websites, and we used three large language models (LLMs)—Gemini, ChatGPT-4o, and Claude 3.5 Sonnet—as evaluators in the verification process. Only samples that received two or more votes for correctness were considered. four techniques were evaluated: zero-shots, few-shots, zero-shots chain of thought (COT) and COT

1. Zero-shots:

Zero-shots is the basic one which is simply involving asking the model if the I'rab correct or not. And if it is not correct, write the correct one to help players to learn.

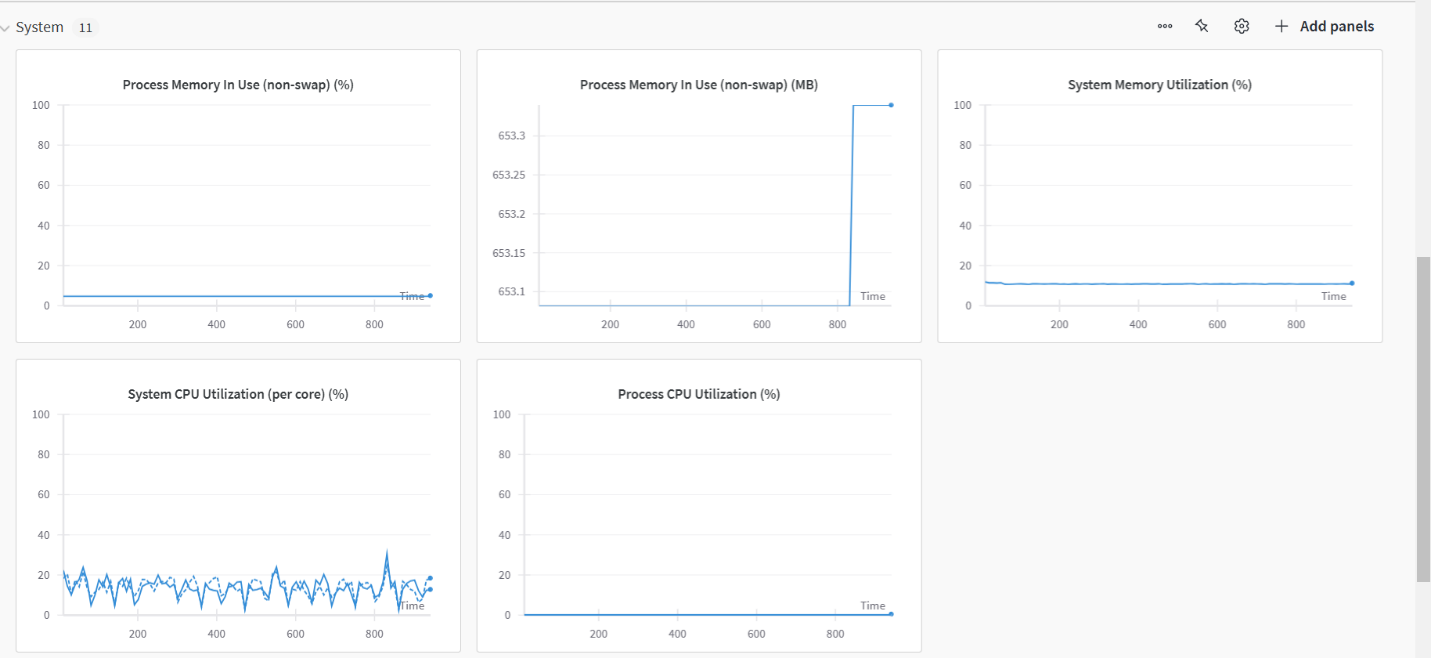
**Performance metrics:**

Accuracy: 86.96 %

Average response time: 1.27 seconds

A graph with blue lines

Description automatically generated



A screenshot of a graph

Description automatically generated

1. Few-shots:

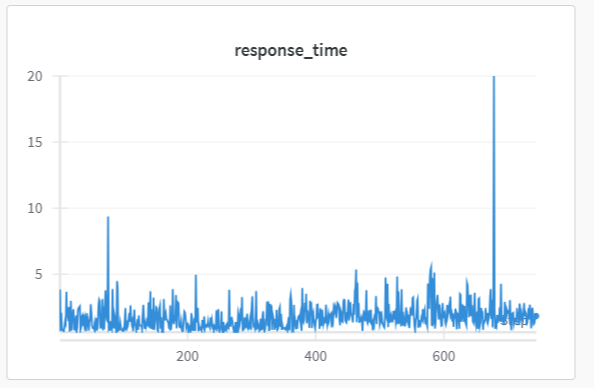
We included 10 examples (shots) as part of the input prompt to guide the model's responses and improve accuracy

**Performance metrics:**

Accuracy: 95.46 %

Average response time: 1.83 seconds

Charts:



A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

1. COT:

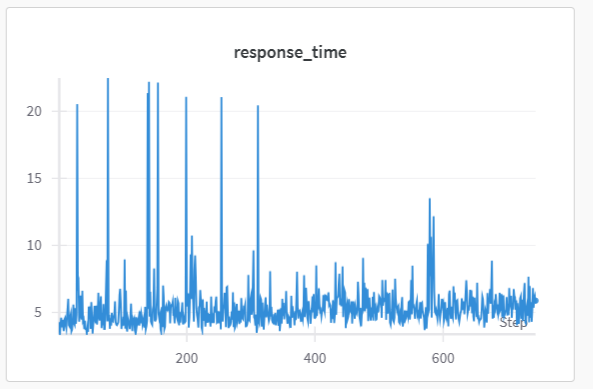
We applied 10 examples using step-by-step reasoning to help the model produce more logical and detailed answers

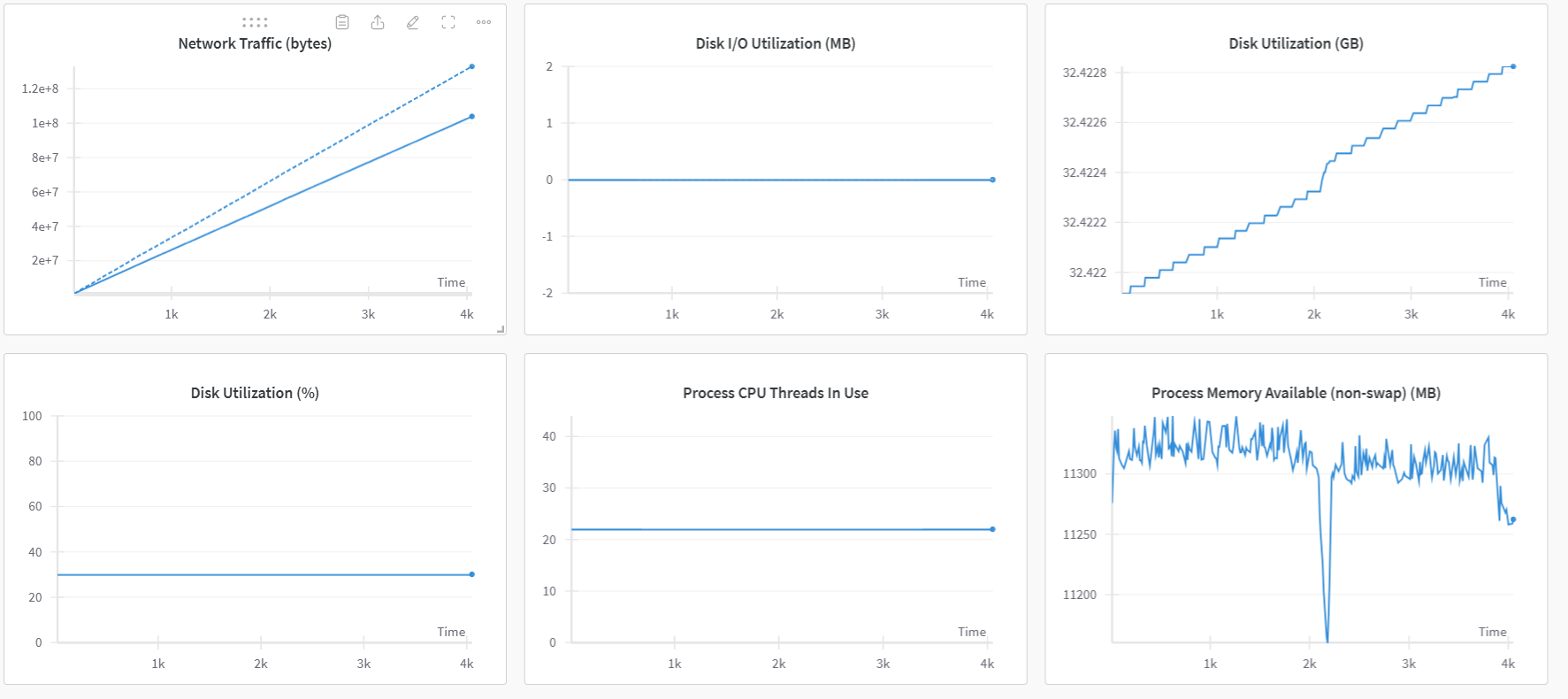
**Performance metrics:**

Accuracy: 92.61%

Average response time: 5.45 seconds

Charts:





A screenshot of a graph

Description automatically generated

1. Zero-COT:

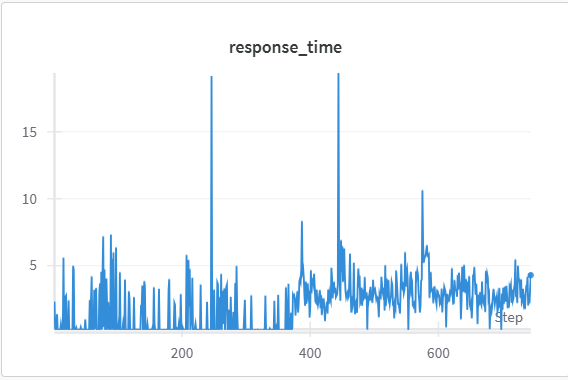
Same as Zero-shots, but we include this phrase at the end of each input prompt " لنقم بتحليل الجملة خطوة بخطوة:"

**Performance metrics:**

Accuracy: 88.44%

Average response time: 2.01 seconds

Charts:

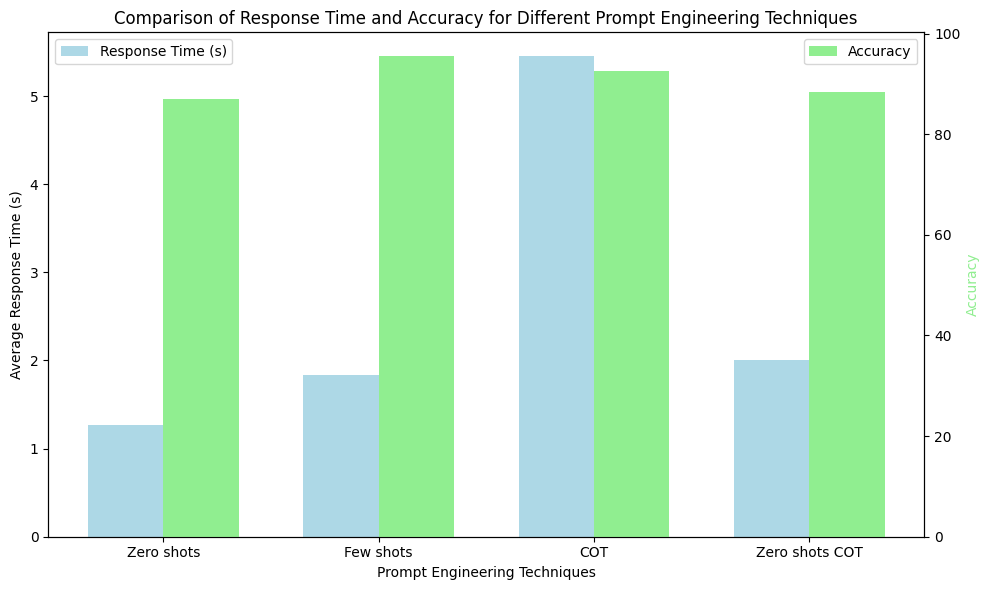




A screenshot of a graph

Description automatically generated

Summary chart:



Conclusion:

In conclusion, we found that few-shot prompting provides the best balance of accuracy and response time, closely approaching the speed of zero-shot methods while significantly enhancing performance. Given this advantage, we have decided to proceed with few-shot prompts for optimal results.