NLP PROJECT: IF YOU ASK NICELY…

Sena TOKER

1. **INTRODUCTION**

Prompt engineering has emerged as a powerful tool in shaping the output of large language models. The way a task is formulated the way a question asked to a large language model influences the outcome of the model, especially when facts are less crucial than the creativity. This project investigates how different types of prompt types affect the generation of both open and closed source models in terms of coherence, creativity, adherence to prompt instructions, style, and hidden elements while also focusing on underlying confidence through log probability analysis.

This project focuses on a specific task: providing a scientifically inaccurate yet plausible explanation of how black holes enable time travel. While numerous works have explored prompt techniques for factual QA, summarization, and reasoning, fewer studies have examined their effects in creative settings.

Five distinct prompt styles employed: Explicit Instruction, Few-Shot, Emotionally Stimulated, Role-based, and Zero-Shot. Then, completions generated from both GPT-4 (closed) and Mistral-7B-Instruct (open), followed by comparative evaluation through both qualitative and quantitave methods.

1. **RESEARCH QUESTION AND METHODOLOGY**

**Research Question**

How does prompt framing influence the output quality and token-level confidence (log probability) of large language models in creative generation tasks?

**Objectives**

* To evaluate the stylistic and structural differences in outputs generated under five prompting strategies.
* To analyze the log probability values as a proxy for model confidence.
* To analyze prompt attribution to explore which prompt sections influenced the generation most.
* To compare and open-source model (Mistral) against a closed source (GPT-4).

**Prompt Design**

1. **Explicit:** Direct instructions.
2. **Few-Shot:** Examples are given.
3. **Emotionally Stimulated:** Framing with emotional tones.
4. **Role-based:** Adopting a character (science Youtuber)
5. **Zero-Shot:** Under-specified prompt.

**Output Generation**

* GPT-4, OpenAI Playground (temperature = 0.7, max tokens = 300). Since API key access was restricted during experimentation, outputs were obtained externally from a trusted assistant. (GPT-4)
* Mistral-7B-Instruct-v.01, using Hugging Face’s Transformers pipeline. Most generations were successfully obtained; however, the few-shot completion and subsequent explainability analyses were conducted externally due to repeated execution errors.

**Explainability Methods**

* **Log Probability Analysis:** Extracting logprobs to measure generation confidence.
* **Prompt Attribution:** Using a modified forward pass to correlate prompt tokens with output logits.

1. **EXPERIMENTAL RESULTS**
   1. **Output Evaluation Metrics**

Each model output was evaluated based on five dimensions:

* **Coherence:** Logical consistency
* **Creativity:** Use of metaphor or imagination
* **Adherence to Prompt:** Degree to which the model fulfills the goal.
* **Style:** Tone, voice and rhetorical richness.
* **Hidden Elements:** Presence of subtext, metaphors, or layered meaning.

Scores were assigned on a scale from 1 to 5.

**3.2 Prompt Type Performance Comparison**

A table of style evaluation

Description automatically generated with medium confidence

A graph of different colored bars

Description automatically generated

**Observations**

* Emotionally stimulated prompts achieved the highest overall performance.
* Few-Shot prompting also performed well across both models.
* Zero-Shot prompting performed the weakest, especially for GPT-4.
* Explicit instructions yielded average results.
* Role-based prompting yielded a gap between models.
* GPT-4 slightly outperformed Mistral in most prompt categories.

**3.3 Explainability Analysis**

To explore how different prompt styles affected model behaviour beyond surface-level output two explainability methods are applied: log probability analysis, and prompt attribution.

**3.3.1 Log Probability Analysis**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**Observations**

* Both models showed a pattern: creativity came at the cost of lower logprobs, especially for Few-Shot and Emotionally Stimulated Prompts.
* Higher logic deltas were seen in more rigid prompts (Explicit, Role-based), implying less uncertainty.

**3.3.2 Prompt Attribution**

Prompt attribution aims to determine which parts of a prompt strongly influenced the output, token by token. For each output token, each prompt token contribution to its generation is measured.

**Key Observations**

* Few-Shot prompts consistently showed high attribution to the in context examples.
* Emotionally Stimulated prompts showed strong influence from emotional phrases like “mind blowing secret”, which heavily shaped tone and style.
* Explicit prompts relied most on directive keywords like “explain” , leading to less creative responses.
* Role-based prompts had mixed attribution. Character distribution (You are a Youtuber) and the task.
* Zero-Shot prompts had the most dispersed attribution, weak concentration to specific prompt parts.

**Sample Visualization**

A graph with orange and yellow bars

Description automatically generated

1. **CONCLUDING REMARKS**

This study explored how prompt design shapes the behaviour of large language models in a creative setting. Experiments revealed that emotionally stimulated and few-shot prompts enhance both output quality and style, as reflected in human ratings and explainability metrics. Models showed higher uncertainty when creativity was encouraged, suggesting a trade-off between creativity and confidence.

While findings shows insights into prompt engineering, further work could explore:

* Applying prompt attribution across diverse genres and tasks.
* Exploring prompt refinement in real time using attribution feedback.

**ACKNOWLEDGEMENTS**

Throughout the completion of this project, I utilized GPT-4 as a supportive tool for organizing report sections, assisting with code debugging, and generating visualizations based on analysis results. While the conceptual design, prompt development, and critical evaluation were conducted independently, GPT-4 was instrumental in enhancing the clarity and structure of the final output.