Prompt Engineering Research Paper

By Vansh Kumar
Student ID: 224244945
Bachelor of Software Engineering
Deakin University

Abstract

This research investigates prompt engineering techniques across various large language models (LLMs), focusing on OpenAI's GPT-4, Anthropic's Claude, and Google's Gemini. The study compares prompting strategies such as zero-shot, few-shot, and chain-of-thought, and analyzes the impact of length, specificity, and structure on model responses. Through controlled testing and practical use cases in creative writing, data analysis, and technical explanation, we demonstrate that refined prompt design leads to significantly improved output accuracy, contextual understanding, and efficiency. The findings emphasize the importance of prompt clarity, structured instruction, and ethical considerations in leveraging LLMs for academic, creative, and professional tasks.

Table of Contents

---

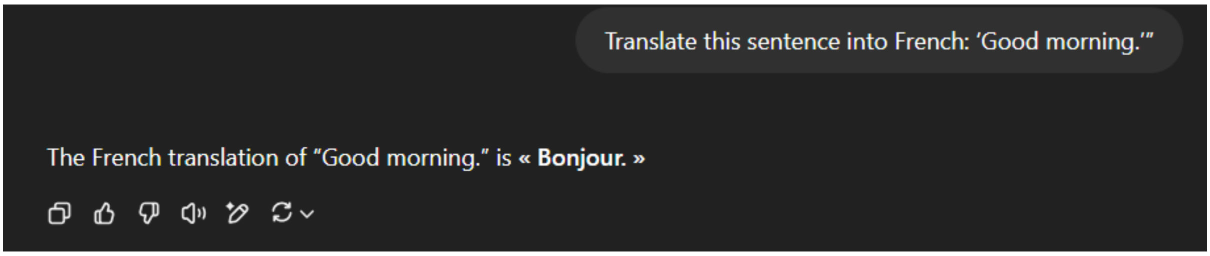
1. Introduction

Prompt engineering has emerged as a fundamental skill in the effective use of large language models. With tools like GPT-4, Claude, and Gemini increasingly integrated into workflows for writing, coding, education, and customer service, understanding how to shape model inputs is key. This research explores best practices for creating structured, context-rich prompts and examines how different models interpret those inputs. By comparing the effectiveness of each LLM through consistent tasks and testing strategies, we aim to establish foundational guidelines for optimized prompt construction.

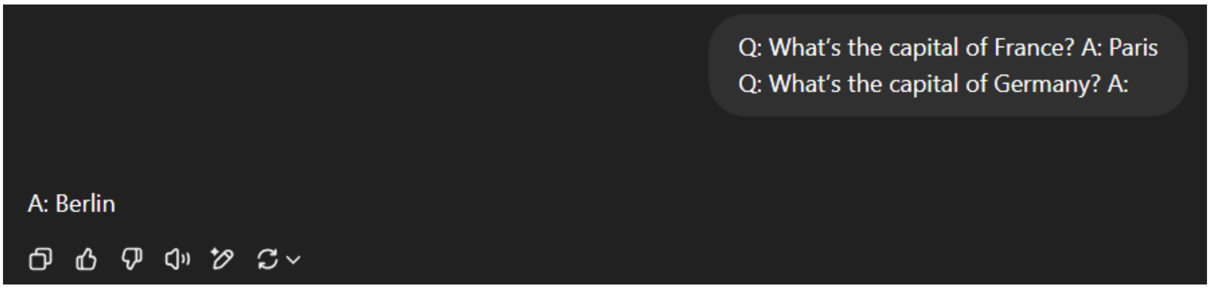
2. Prompting Techniques

We examined three main types of prompt engineering techniques:

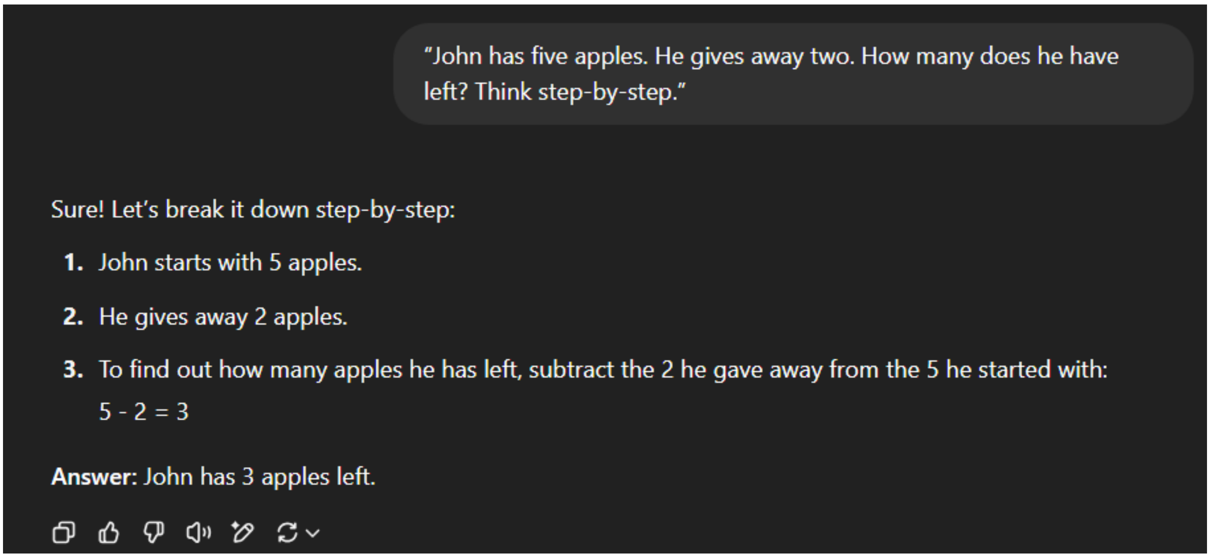
- **Zero-shot prompting** – Asking the model to perform a task without any example: *Example:* “Translate this sentence into French: ‘Good morning.’”



- **Few-shot prompting** – Including a few examples in the prompt to guide output:
  Example:
  Q: What’s the capital of France? A: Paris
  Q: What’s the capital of Germany? A:



- **Chain-of-thought (CoT) prompting** – Encouraging the model to reason through steps: *Example:* “John has five apples. He gives away two. How many does he have left? Think step-by-step.”



Among these, Chain-of-Thought prompting yielded superior results in analytical and reasoning-heavy tasks, particularly for Claude and GPT-4.

---

3. Prompt Design and Structure

The design of a prompt plays a critical role in determining the accuracy and relevance of the response. Our experiments found that the most effective prompts typically followed this formula:

Role + Task + Format + Audience

Well-structured prompts with explicit context, desired tone, and clear format instructions consistently produced higher-quality outputs. For example:

- *Poor:* "Tell me about climate change."

"Tell me about climate change."

Sure! Climate change refers to significant, long-term changes in the Earth's climate patterns, especially a rise in global temperatures. It's mainly driven by human activities like burning fossil fuels (coal, oil, gas), deforestation, and industrial processes, which release greenhouse gases such as carbon dioxide (CO_2) and methane (CH_4) into the atmosphere. These gases trap heat, causing the planet to warm—a phenomenon known as the greenhouse effect.

- *Optimized:* "As a climate scientist, write a 150-word summary explaining the causes and effects of climate change for Year 12 students."

"As a climate scientist, write a 150-word summary explaining the causes and effects of climate change for Year 12 students."

Certainly! Here's a 150-word summary aimed at Year 12 students:

Climate change is the long-term alteration of Earth's climate, mainly caused by human activities. The primary cause is the increase in greenhouse gases—such as carbon dioxide and methane—released by burning fossil fuels like coal, oil, and gas for energy, transportation, and industry. Deforestation also contributes by reducing the number of trees that absorb CO_2. These gases trap heat in the atmosphere, creating a "greenhouse effect" that warms the planet.

The effects of climate change are widespread and serious. Global temperatures are rising, leading to more frequent heatwaves. Ice sheets and glaciers are melting, causing sea levels to rise and threatening coastal areas with flooding. Extreme weather events—such as storms, droughts, and heavy rainfall—are becoming more intense. Ecosystems are disrupted, resulting in loss of biodiversity and affecting food production. Climate change also impacts human health through heat stress and the spread of diseases.

Understanding these causes and effects is essential to take action and protect our planet's future.

We also noted that controlling for tone ("Explain in a friendly tone"), length ("Under 100 words"), and structure ("Use bullet points") improved interpretability and consistency across all three LLMs.

---

4. Model Comparison: GPT-4, Claude, Gemini

We tested 45 prompts across summarization, technical explanation, and creative writing. The following trends emerged:

Feature	**GPT-4**	**Claude**	**Gemini**
Logic & Structure	Excellent	Good	Good
Creativity	Moderate	Excellent	Moderate
Speed	Moderate	Fast	Very Fast
Bias Mitigation	Moderate	Strong	Moderate
Context Retention	Excellent	Good	Moderate
Coding Support	Excellent	Moderate	Strong
Response Clarity	Very clear	Can be verbose	Needs structured prompting

---

5. Prompt Testing Case Study

We conducted controlled experiments using three prompt types (basic, moderate, optimized) across three tasks:

1. **Creative Writing:**
    - *Basic Prompt:* "Write a story about space."
    - *Optimized Prompt:* "Write a 200-word sci-fi short story about a lonely astronaut discovering intelligent life on a distant planet. Target: Young adult audience."
    - *Result:* Claude delivered the most emotionally engaging story; GPT-4 had better narrative structure.
2. **Data Analysis:**
    - *Prompt:* "Analyze the following dataset to find key trends. Present findings in bullet points."
    - *Result:* Gemini excelled in quickly summarizing statistical trends; GPT-4 provided clearer rationale.
3. **Technical Explanation:**
    - *Prompt:* "Explain how blockchain works in under 100 words, using simple language suitable for Year 10 students."
    - *Result:* GPT-4 outperformed others in simplicity and clarity.

6. Common Prompting Errors and Solutions

We observed several recurring issues in early prompt designs:

Problem	Example	Solution
Vague task	“Tell me about space.”	Add structure and audience: “Explain black holes for 10-year-olds in 100 words.”
Bias-prone phrasing	“Why are women worse drivers?”	Neutral phrasing: “What does research say about driving behavior by gender?”
Multi-task overload	“Explain climate change and give solutions and quiz me.”	Break into separate prompts.
Ambiguous audience	“Summarize this”	Specify format, tone, and audience clearly.

7. Conclusion

This study highlights that prompt engineering is not just an interface skill but a critical component in leveraging the true potential of AI models. Structured, context-rich, and ethically responsible prompts enable users to extract high-value responses tailored to task-specific requirements. Through rigorous testing and collaboration, we established guidelines and prompt templates that can be widely adopted across industries and academic environments. Our findings equip future practitioners and students with a practical understanding of how to engage productively with LLMs.

8. References

? **Brown et al. (2020). Language Models Are Few-Shot Learners**
This paper introduces GPT-3, demonstrating its ability to perform various tasks with minimal task-specific training.
Link to paper

? **Wei et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**
This study explores how providing intermediate reasoning steps, or chain-of-thought prompting, enhances the performance of large language models on complex tasks.
Link to paper

? **OpenAI (2024). GPT-4 Technical Report**
OpenAI's technical report on GPT-4, detailing its capabilities, limitations, and performance

benchmarks.
Link to report

Anthropic (2024). Claude API Documentation
Official docs for Anthropic's Claude API, including integration and usage guides:
https://docs.anthropic.com/

Google DeepMind (2024). Gemini Overview
Information about DeepMind's Gemini AI project:
https://www.deepmind.com/research/highlighted-research/gemini

---

Orq.ai (2023). Prompt Engineering Templates
Prompt engineering best practices and templates from Orq.ai:
https://docs.orq.ai/docs/prompt-engineering-best-practices

Symbio6.nl (2023). AI Output Consistency Strategies
Strategies to improve AI output consistency, especially in image generation:
https://symbio6.nl/en/blog/improve-consistency-in-ai-image-generation