SE Proj1d1 - Group 19

Digvijay Sanjeev Sonvane
Suyesh Jadhav
Vanaja Agarwal

Reflection on LLM Prompting Experiments

Pain Points in Using LLMs

The major challenge was the high discard rate of outputs, with the majority of the responses proving overly generic or superficial. This inefficiency delayed progress, since useful results often required several rounds of refinement. Another issue was hallucination and drift: as conversations extended, models increasingly deviated from the original scope, introducing features or details never specified in the requirements. Maintaining consistency across sessions proved difficult, exposing the fragility of relying on unstructured dialogue.

Surprises

An unexpected finding was the difference in how models structured their outputs. One favored technical jargon and system-design terminology, while another prioritized clarity and accessibility for non-technical stakeholders. This divergence demonstrated that the same prompt could generate outputs tailored to very different audiences, which, though surprising, was helpful for realizing that model selection inherently influences communication style. Conclusions also sometimes varied: one model emphasized compliance-heavy processes, while the other focused on health, sustainability, and user experience.

What Worked Best

The most effective technique was LLM-to-LLM prompting, where one model generated or refined prompts and another produced the final output. This “two-step pipeline” reduced genericity and improved structure, fostering a feedback loop that drove outputs to higher standards. Additionally, Chain-of-Thought prompting proved invaluable: by requiring stepwise reasoning, we consistently obtained deeper and more logically coherent responses, rather than superficial lists.

What Worked Worst

The least effective approach was zero-shot prompting. Without examples or constraints, outputs were vague, repetitive, and often wrong. It was particularly vulnerable to hallucinations, especially when prompts were broad or when conversation history became lengthy. Similarly, extended uncurated sessions degraded quality, as the model increasingly contradicted itself or invented details to fill perceived gaps.

Pre- and Post-Processing Value

The most useful pre-processing involved structuring prompts with explicit examples, success criteria, and constraints. Instead of leaving interpretation open, prompts that defined the format and scope produced consistently stronger results. Post-processing was equally critical: having a second model summarize, validate, or simplify the first model's output revealed gaps and reduced errors. This layered workflow prevented drift and allowed humans to focus on higher-order evaluation rather than basic correction.

Best and Worst Prompting Strategies

The best strategies were:

- LLM-to-LLM prompting for refinement and validation.
- Chain-of-Thought prompting to enforce reasoning depth.
- Example-driven prompts to guarantee structure and reduce ambiguity.

The worst strategies were:

- Zero-shot prompting, which failed to provide reliable or nuanced responses.
- Long, unscaffolded conversations, which led to hallucinations and off-scope content.