Project 1d1 (Group 2)

Jawad Saeed (jsaeed@ncsu.edu)
Mason Cormany (mcorman@ncsu.edu)
Himanshu Agarwal (hagarwa4@ncsu.edu)
Omkar Joshi (ojjoshi@ncsu.edu)
Mulya Patel (mspate22@ncsu.edu)

When reflecting upon the best way to use LLMs for planning the university food delivery software, these are some common thoughts from Group 2.

The pain points in using LLMs largely depended on the specific model in question. For instance, with Claude, one recurring issue was the length of the context window. As we worked through different prompts, especially when trying to generate responses from the perspectives of various stakeholders, we quickly hit the model's context window limit. This forced us to start new chats and reintroduce the entire context each time, which significantly reduced the effectiveness of the conversation.

With ChatGPT, the challenges were more related to prompt structure. We found that well-structured prompts yielded far better results than raw, unformatted input. Additionally, poor instruction-following, particularly with tools like Deep Research, presented issues.

A major surprise came from comparing the results between Claude and ChatGPT. Claude excelled at generating content that stayed true to the overall goal and offered more creative outputs, such as identifying regulators (e.g., FDA) as stakeholders and proposing unique use cases like inspections and booking appointments. In contrast, ChatGPT was better at adhering to formatting instructions, whether in zero-shot or multi-shot prompting. This distinction in their strengths was unexpected but insightful, highlighting the nuances in model design and capabilities.

The most successful approach for us was multi-shot prompting, which consistently yielded good results across projects. It allowed us to break down tasks into manageable chunks and refine the output over multiple iterations. For initial use case generation, Persona Generation turned out to be incredibly effective, providing granular insights into use cases, biases, and even highlighting potential conflicts between stakeholders—something that other methods failed to capture as effectively.

One standout implementation was our chain-of-thought approach for Project 1c1. While simple, it produced consistent and well-defined Minimum Viable Products (MVPs) across models, and the prompt chain we used gave us detailed feedback at each stage of the process. Additionally, most models we worked with performed well when it came to formatting, providing structured output that was easy to process and work with.

Zero-shot prompting was notably less effective, primarily generating generic use cases with little depth. This limitation became especially apparent when compared to other strategies like Chain-of-Thought or Persona Generation, which provided much more detailed and actionable results. We also experimented with a few free APIs for RAG (Retrieval-Augmented Generation), such as DeepSeek and Gemini. These models produced solid results for generating a small number of detailed use cases, but as soon as we scaled up the number of use cases, the quality significantly dropped.

A key aspect of success was the pre-processing of prompts. Instead of throwing everything into a single, overwhelming prompt, we broke it down into sections—first providing context, then the actual instruction, and finally specifying the desired format. Templates were extremely helpful in guiding the models toward structured outputs, making them easier to work with. Examples were also invaluable in helping the model stay on track and focus on the critical information we needed. This helped with post-processing and separating the weaker results from the stronger ones.

In the end, some clear cut best and worst strategies were found. These best strategies included multi-shot prompting, which offered consistent and high quality results, and persona generation. The worst strategies found were zero-shot prompting, which gave very generic and shallow answers, and using free APIs, as they lacked detail and did not hold up well in relation to other strategies.