Course Project Proposal
Institution Profiling using LLMs and Prompt Engineering

Noura Manassra – 1212359 —— Mariam Turk – 1211115 —— Abdallah AboShoaib -1210211
Mohammad Salama - 1212109

May 1, 2025

1. Motivation and Background

Large Language Models (LLMs) such as OpenAI's GPT-4 have demonstrated exceptional capabilities in processing and generating natural language text. However, using these models effectively for specific tasks such as structured data extraction and summarization still requires thoughtful prompt engineering[2] and contextual data feeding.

In today's data-driven world, institutions like banks, hospitals, and universities maintain complex, dynamic profiles that are important for various stakeholders, including customers, investors, students, and researchers. Yet, much of this information is scattered across websites, press releases, and reports, making it difficult to obtain a structured overview quickly.

Our motivation for this project is to leverage the power of LLMs, combined with carefully engineered prompts and selective data collection, to automate the creation of concise, informative institutional profiles. By doing so, we aim to demonstrate practical applications of modern NLP tools and reinforce our understanding of real-world language processing techniques.

This project aligns with the course's objectives by applying NLP principles, data retrieval methods, and API integration, while also offering us a hands-on opportunity to explore advanced prompt engineering strategies.

2. Project Management Plan

Our project will progress through four main phases:

1. **Prompt Design:** We will experiment with various prompt structures to guide the LLM towards outputting structured, high-quality institutional profiles. We will refine prompts based on early results to improve consistency and reliability.
2. **Data Collection:** Instead of scraping arbitrary web content, we will target reliable sources such as official institution websites, Wikipedia, and reputable news portals. This ensures the LLM operates on clean and factual data inputs.
3. **System Implementation:** We will build a Python-based interface that manages the data collection, formulates dynamic prompts, interacts with the OpenAI API[1], and processes the outputs into final structured profiles. This will include modules for preprocessing input data and post-processing LLM outputs.
4. **Evaluation and Refinement:** We will create a small benchmark dataset by manually assembling correct profiles for a few institutions. The generated profiles will be compared against these benchmarks to assess coverage, factual correctness, and presentation quality. User feedback will also be collected for qualitative evaluation.

3. Expected Outcomes

By the end of the semester, we expect to deliver:

- A working system capable of generating structured institutional profiles from a given institution name.
- A comprehensive project report documenting the methodologies, challenges, and results.
- A presentation demonstrating the system's functionality and findings.

4. References

1. OpenAI. OpenAI API Documentation. Available at: `https://platform.openai.com/docs`
2. dair-ai. Prompt Engineering Guide. Available at: `https://github.com/dair-ai/Prompt-Engineering-Guide`