Report

HW1

The main goal of the project, as identified as Homework 1, is to build an Agentic AI system that is capable of extracting specific types of financial information from the images of the receipts. In the context of modern data processing systems, the ability to automate the extraction of structured data from unstructured inputs is a crucial one. The current project is focused on the use of Large Language Models (LLMs) for the interpretation of the visual data, specifically through the Google GenAI interface. The system is not simply meant to perform Optical Character Recognition (OCR) tasks but also needs to be able to interpret the context of the data presented in the receipts so that it can answer complex natural language queries such as the total amount spent or the original prices before any discount. Through the integration of the LangChain framework with the Google generative models, the project aims to build a strong system that is capable of handling image inputs and providing accurate text-based output.

The core methodology is based on the integration of the langchain-google-genai library, which acts as a bridge for the Python execution environment and the powerful Gemini model from Google. The development environment has been set up using a Jupyter Notebook, which provides a modular approach to coding. The first step in the development process has been the setup of the required dependencies and the API credentials to enable authentication with the Vertex AI service. This is a vital step in enabling the transmission of image data and instructions to the model. The architecture is based on a "Chain" concept, which is a fundamental component of LangChain. It is responsible for the sequencing of the prompt and the image data in a logical request format. The architecture has been designed to process multimodal input, where the receipt images are not simply files but informational contexts from which the AI model needs to derive the correct numerical values.

Another important part of the implementation process was the creation of the prompting strategy and the logical flow for the data extraction process. This is achieved through the function get_receipt_data_with_chain, which acts as the main functional unit. It contains the code needed to send the image as well as the text query to the model. To ensure the accuracy of the response, the code implements different queries. For basic queries, such as the total amount spent, the model is asked to scan the visual parts of the image, find the "Total" field, and provide the corresponding information. More complex queries, such as those for the prices "without discount," require the agent to imply the need for finding the subtotals of all items and summing them, or finding a "subtotal" field before any coupons are applied. This shows the agent's ability to perform arithmetic calculations based on the visual information, rather than the actual text. The code goes through a list of images provided for the receipt, applying the logical chains sequentially to test the prompts for different types of receipts.

In order to further elevate the system from its status as merely an extractor to one that is much more robust and agent-like, a specific mechanism for handling irrelevant queries was implemented. This is an extremely important component of any AI system, as the queries that users input into the system can be completely unpredictable. A specific test case, referred to as "Query 3," was implemented to test the rejection mechanism. In the prompt, the model was given specific instructions that, if the query did not pertain to the extraction of information on the receipt, the model should output the strict token "IRRELEVANT" and refuse to answer. The execution logs of the notebook confirm that the rejection mechanism was successful. When the system was given prompts that were not relevant to the receipt, the chain correctly recognized the disconnect between the query and the available data context. Rather than responding to the query, the model, following the safety constraints, outputted "IRRELEVANT." This is clear evidence that the prompts given to the system were successful in constraining the model's output, ensuring that the agent is only focused on the specific domain of receipt analysis.

In conclusion, this project was successfully implemented to prove the viability of Agentic AI workflows in automating financial document processing. Through the use of the langchain-google-genai library, the system was able to successfully process several receipt images, responding appropriately to queries regarding total cost and pre-discount values. The successful implementation of the rejection logic also speaks to the maturity of the system, demonstrating the ability of the system to distinguish between valid and invalid queries. The results obtained from the notebook execution demonstrate that Large Language Models are multimodal enough to replace the traditional approach to OCR and Regex. The future of this work could be to scale this approach to automate the processing of thousands of receipts or to integrate it with other accounting software. As it stands, this project meets the requirements for Homework 1, demonstrating a functioning intelligent agent that is able to comprehend visual financial data.