Large Language Models in automated tabular data analysis

Authors

Filip Kołodziejczyk
Jakub Świstak

Supervisor

prof. dr hab. inż. Przemysław Biecek

The growing need for advanced data analysis in decision-making has led to exploring Large Language Models (LLMs) as a potential solution. Current LLMs, however, cannot autonomously conduct analyses. Constant user attention is required. A multi-agent LLM system, combining two LLMs, a theoretical data analyst and a programmer, is proposed to overcome this. This approach aims to enhance autonomous data analysis and coding capabilities, minimizing the user's effort. This proposition was implemented as a Minimum Viable Product (MVP) and tested, focusing on the effectiveness of the LLMs and prompts used.

1 Realized Objectives

Literature Review

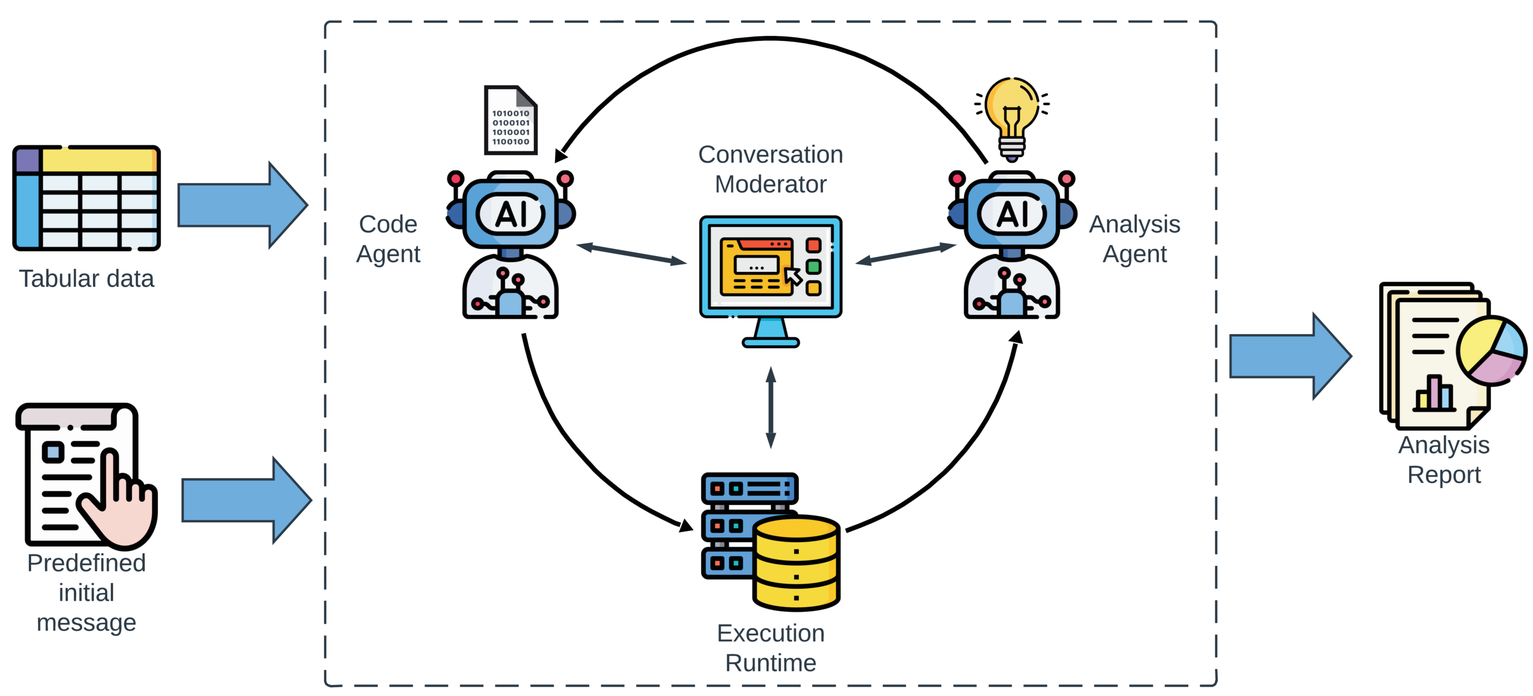
Solution Design

Real-Life Implementation

Experiments

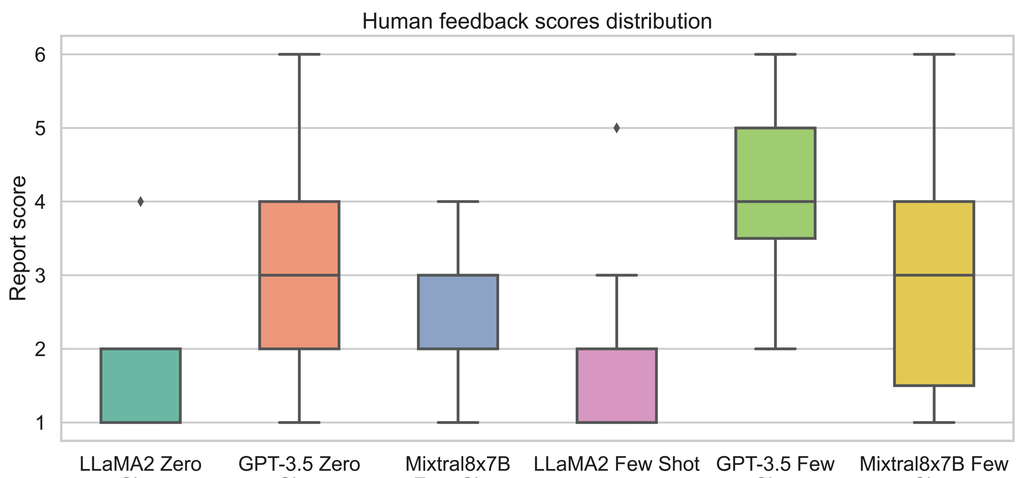
2 Proposed Solution

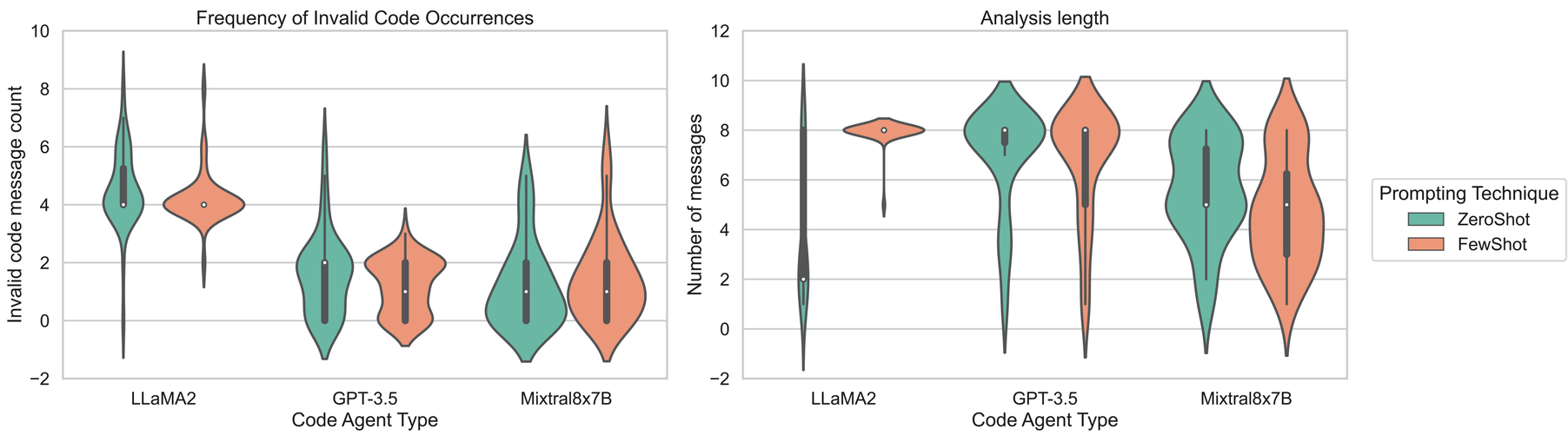
The system comprises two LLM Agents engaged in an infinite dialogue loop. The Analysis Agent summarizes previous code outputs and generates the next steps for data analysis, while the Code Agent transforms these steps into executable Python code. The roles of these agents are defined and maintained by specific system prompts. To ensure security, an execution runtime environment is used for code operation. A moderator oversees all internal conversation intricacies and error handling, ensuring smooth operation. The only user's input is the dataset. A hardcoded Python script that provides an overview of the dataset initiates the cycle. The system's output, the analysis report, consists of messages from the Analysis Agent, the Python code, and its execution results.



3 Experiments

The system underwent two types of tests to assess its effectiveness: Code Agent Output Quality and Human Feedback. The Code Agent Output Quality test involved a simulation with a single dataset, generating multiple reports for each parameter configuration and counting the number of invalid outputs, which included messages without code snippets or with snippets causing errors. In the Human Feedback test, volunteers provided their datasets, for which reports were generated based on each parameter configuration. These reports were then graded by the volunteers on a scale of 1 to 6, providing valuable insights into the quality of the reports from the users' perspective.





Faculty of Mathematics and Information Sciences
WARSAW UNIVERSITY OF TECHNOLOGY

https://github.com/jswistak/automatic-data-analysis