**Design Document Hackathon 2025**

**Gen Ai-powered Data Profiling**

**Elite-V**

Jyoti R Swain

Kartik Ramanathan

Likith S

Dhanya H U

Sachin Bontadka

**Contents**

[**Problem Statement:** 4](#_Toc161586621)

[**Solution Approach:** 5](#_Toc161586622)

[**Workflow:** 6](#_Toc161586623)

[**Low level architecture** 7](#_Toc161586624)

[**Snapshots and scripts:** 10](#_Toc161586625)

[**Tools used: Software, libraries and Cloud services** 14](#_Toc161586626)

[**Conclusion** 16](#_Toc161586628)

# 

# **Problem Statement:**

Regulatory reporting in the banking sector involves compiling vast amounts of data to meet compliance requirements. A critical aspect of this process is data profiling, which ensures that the reported data aligns with regulatory reporting instructions.

Traditionally, this involves manually defining profiling rules based on the underlying data and regulatory requirements. This challenge aims to automate data profiling using Generative Al (LLMs) and unsupervised machine learning techniques. Participants will develop a solution that can generate data profiling rules, perform adaptive risk scoring, and suggest remediation actions based on regulatory reporting instructions.

With the challenge, we aim to develop a “GenAI-Powered Data profiling” system that leverages Large Language Model (LLM) or Machine learning (ML) algorithms that can generate data profiling rules and automatically flag the transactions based on the allowable values in the Regulatory reporting instructions. By developing this system we intend to Flag the transaction based on the federal rules, provide the Remediation actions with the explanation to the auditors. The system will have the capabilities to conduct the rule based optimization on the input data given to the model

The ML model will be generating relevant and meaningful remediation actions and will be exposed as REST API to evaluate the efficiency.

# **Solution Approach:**

**Problem statement Understanding:**

* Identify the key factors for extracting and creating rules based on the regulatory rules available in Federal reserve document
* Create the rules and apply the same to the input transaction csv to get output which will include the flag and remediation action.

**Data Collection and Preprocessing:**

* Prepared the dataset using the given pdf document, loaded the data and created data frame.
* Preprocessed the data.
  + Extracting the text from PDF
  + Conversion of text to lower case
  + Tokenization
  + Lemmatization
  + Removal of stop words

**Feature preparation:**

* Extracted the key features in regulatory reporting instructions document

**Model:**

Using tools like Pandas and Validation functions (ML) we are creating the model and validating the input transaction csv based on the allowable values and flag any rows which do not meet the validation criteria.

For generating validation rules we have used machine learning model to improve the rule generation and data validation process and to get auto-generated validation rules

**Model Deployment:**

* Created the API in Open AI portal.

**Model validation and Testing:**

* Validated the model for multiple sample data sets for get the best output with the flag and the remediation action.
* Tested the deployed model with multiple test cases with different transactional data.

# **Workflow:**

**End User**

**Problem Statement**

**Model Building**

**Rest API**

**Data Collection**

**Data Preprocessing:**

**Creation of Gen Ai-based Data profiling for the regulatory reporting instructions**

CSV output

**Open AI API**

**Model:**

**AI/ML**



**Model evaluation**

**Model Deployment:**

# 

# **Low level architecture**

+-------------------------------------------------+

| Input Regulatory instructions |

| (User Interface) |

+-----------------------+-------------------------+

+-------------------------+

| Request Handling |

| (Extracts user inputs)|

+-----------+-------------+

+-----------------------+

| Data Validation & |

| Parsing |

| (Clean, format data)|

+-----------+-----------+

+-----------------------+

| Preprocessing |

| (Tokenization, |

| remove stopwords, |

| lemmatization) |

+-----------+-----------+

| |

| **(AI/ML Model)** |

| |

| |

+-----------+-----------+

+-----------------------+

| Matching Algorithm |

| (Compute for rules |

| Rules validation) |

+-----------+-----------+

+-----------------------+

| Response Handling |

| (Excel report |

| Flagged transactions |

+-----------------------+

* **Frontend**: This is the user interface where users interact with the system, providing their job preferences.
* **Request Handling**: Handles incoming requests from the frontend, extracting necessary information for processing.
* **Data Validation & Parsing**: Validates incoming data and parses it into a format suitable for further processing.
* **Preprocessing**: Cleans and preprocesses the text data, including tasks such as tokenization, removing stopwords, and lemmatization.
* **ML Model**: Created validation rules to validate the provided input csv and raises flag as part of validation.
* **Response Handling**: Formats and prints in accessible and readable format(csv).

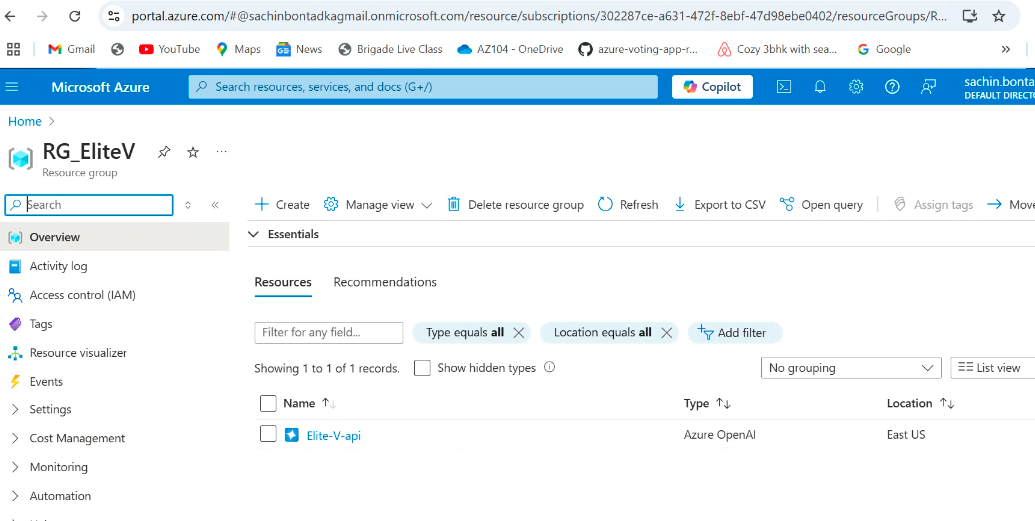
# **Snapshots and scripts:**

**Data collection and readiness:**

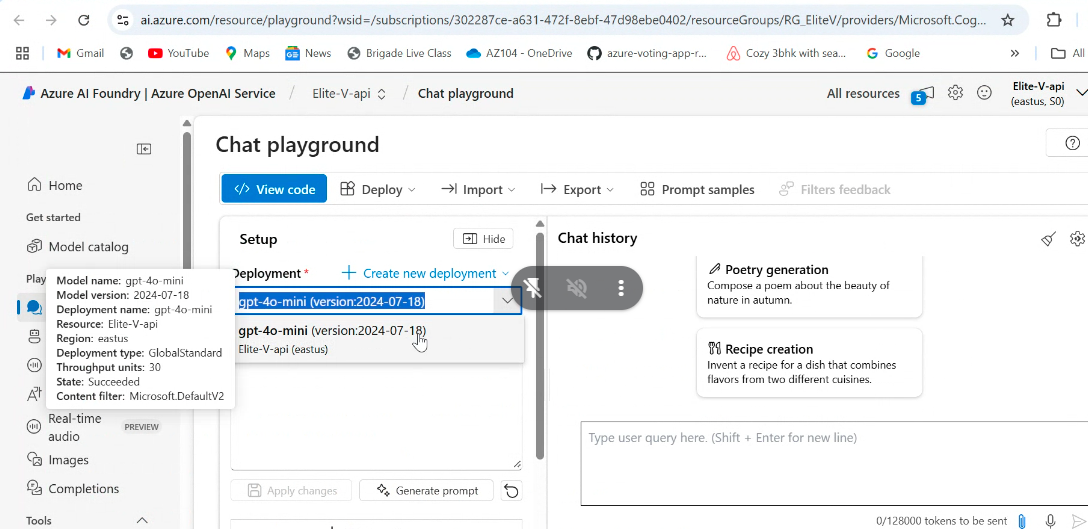
python script for data processing, validation and output:

****

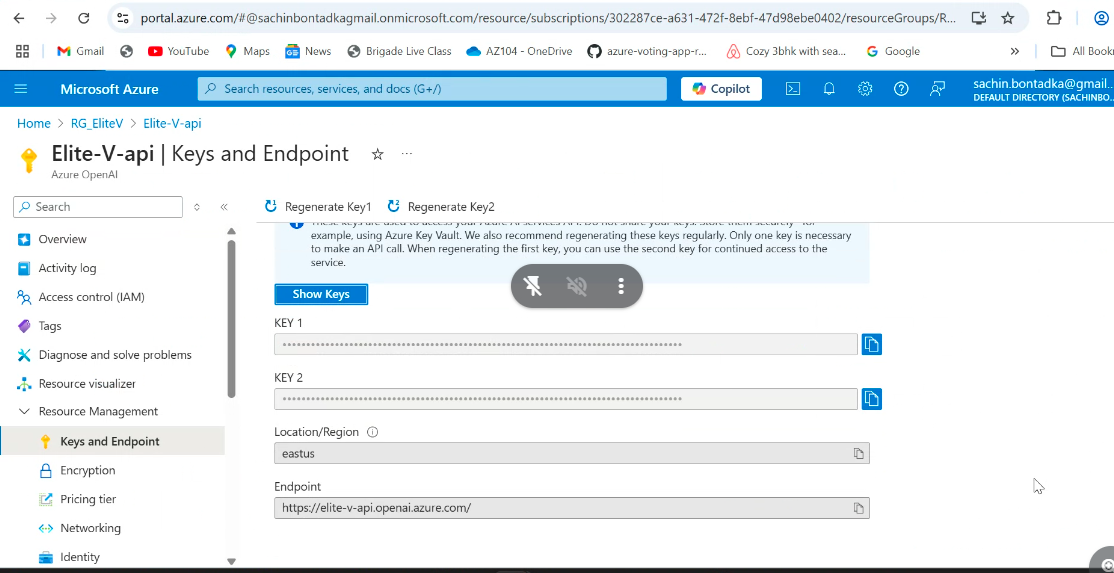
**Bucket:**



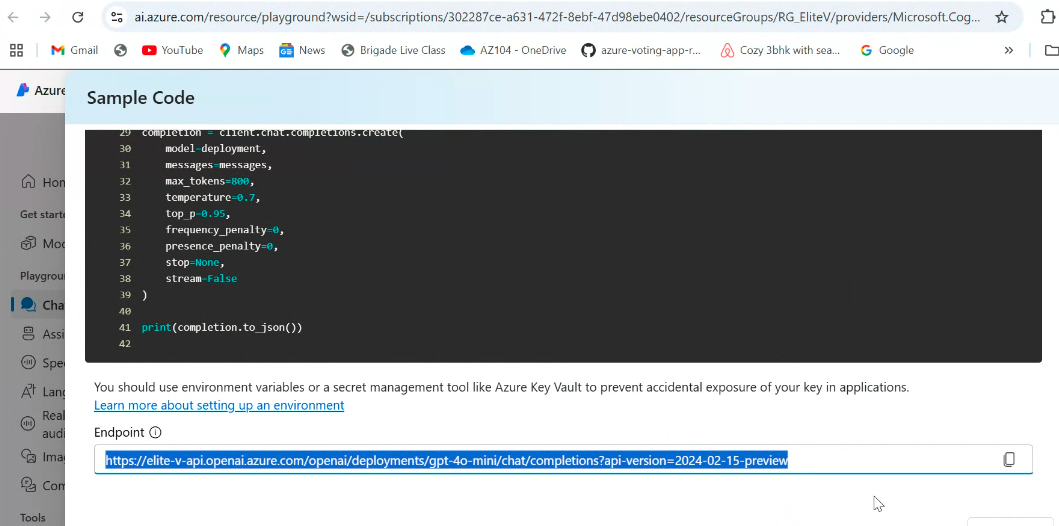
**Model:**



**API Gateway:**



**API endpoint:**



**Results:**

****

# **Tools used: Software, libraries and Cloud services**

Cloud Provider: **Azure cloud**

Cost effectiveness: **Autoscaling based on the requirement the instances are auto deployed**

Platform**: Azure cloud App engine**

Libraries: **pdfplumber**,  **Pandas, NLTK, Docx, openpyxl.**

Programming language: **Python**

LLMs: **OpenAI**

# **Conclusion**

The proposed system leverages AI/ML model to find the best solution for the Data profiling.

With the challenge, we tried to develop a “GenAI-Powered Data profiling” system that leverages Large Language Model (LLM) or Machine learning (ML) algorithms that can generate data profiling rules and automatically flag the transactions based on the allowable values in the Regulatory reporting instructions. By developing this system we intend to Flag the transaction based on the federal rules, provide the Remediation actions with the explanation to the auditors. The system will have the capabilities to conduct the rule based optimization on the input data given to the model

The ML model will be generating relevant and meaningful remediation actions and will be exposed as REST API to evaluate the efficiency.

This design document outlines the key components, implementation details, evaluation plan, and deployment considerations for developing a resume best match system based on AI/ML.