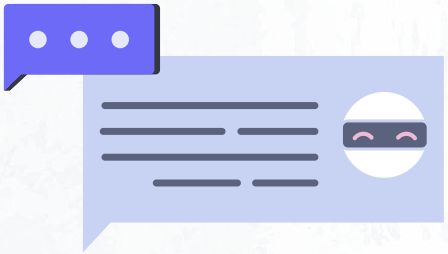
Gen AI-based Data Profiling

Team PHHC

1. Prakhar Agrawal
2. Roshan Prashant Bara
3. Srish Aurangabadkar
4. Sushma Yaramalla



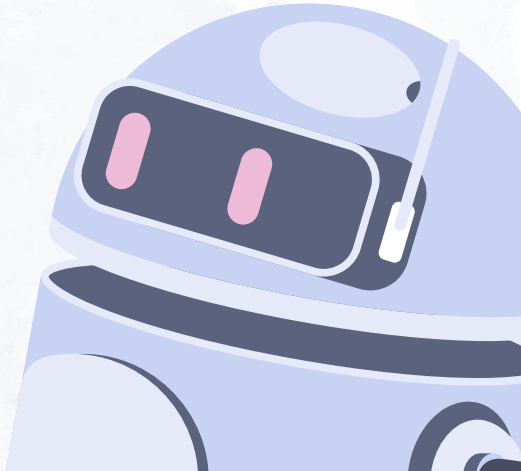


Table of contents

01 →

The Problem:

- Manual compliance checks and validating regulatory reports are slow, lacks efficiency, and difficult to scale.
- Traditional rule generation is time-consuming and prone to human error.

PHHC

02 →

Our Solution

PHHC

Solution Overview:

- A Gen AI-powered data profiling app that automates rule generation from regulatory instructions, converts them into SQL queries, and flags violating transactions
- It ensures fast execution using modern and efficiency libraries
- It provides detailed documentation of the rules generated along with the formulas to validate data

Workflow:

Input Stage:

- Upload regulatory instructions (e.g., policies, compliance guidelines)
- Upload reported data (e.g., transactional records).

Rule Generation with LangChain:

- Gen AI generates profiling rules from regulatory instructions.
- Rules are cached for reuse, reducing repetitive LLM calls.

SQL Query Conversion:

- Rules are converted into SQL queries.

Data Evaluation with DuckDB:

- Queries run on the reported data using DuckDB for fast processing.

Streamlit UI Output:

- Displays flagged transactions with mappings to the violated rules.
- Interactive data filtering and visualization.

03 →

Key Features

PHHC

Key Features

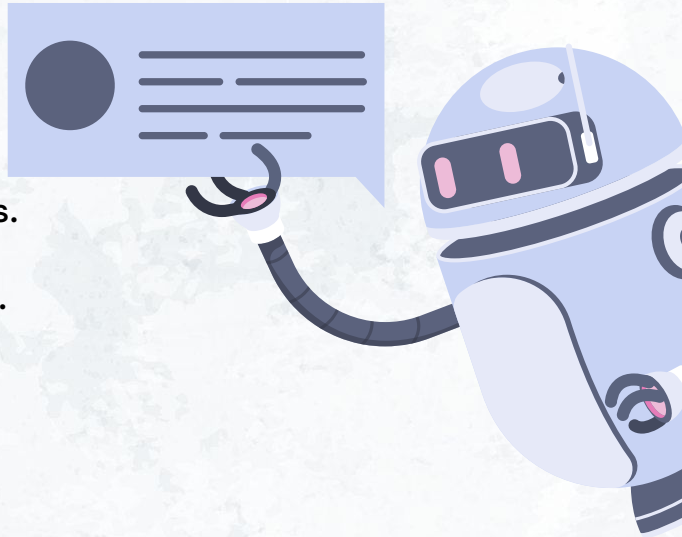
Fast Response Time

- Real-time performance using DuckDB for high-speed SQL execution.
- Efficient processing with minimal latency.

Rule Caching Mechanism

- Caches previously generated rules to avoid repetitive LLM calls.
- Reduces unpredictability from the LLM, ensuring consistent results.
- Enhances performance by lowering API call frequency.

Key Features



LangChain-Powered LLM Integration

- Utilizes LangChain to orchestrate and optimize LLM interactions.
- Efficiently handles rule generation from regulatory instructions.
- Streamlined multi-step prompts for improved rule accuracy.

Streamlit UI for Fast Deployment

- User-friendly UI built with Streamlit for rapid development and deployment.
- Interactive interface for uploading instructions, viewing flagged transactions, and analyzing rule violations.
- Real-time visualization of flagged results.

Key Features

Comprehensive Rule Generation

- Generates profiling rules from all fields in the reported data.
- Creates multiple rules from a single field, covering diverse validation scenarios.
- Improves detection accuracy by applying layered conditions.

Data Security

- No reported data sent to the LLM—only regulatory instructions are processed.
- Ensures sensitive data privacy and compliance with security standards.
- On-premise or controlled processing capabilities.

Key Features

Flagged Transaction Mapping

- Maps flagged transactions to specific rules they violate.
- Provides detailed insights into which regulation each flagged transaction breaches.
- Enhances auditability and simplifies investigations.

04 →

Technical Architecture

PHHC

Technical Architecture

AI Engine:

- LangChain-powered LLM for rule extraction.
- Rule caching layer for efficiency.
- Rule-to-SQL conversion module.

Execution Layer:

- DuckDB for fast, in-memory SQL query execution.

UI Layer:

- Streamlit-based UI for interaction and result visualization.

05 →

Future Scope

PHHC

Future Scope

Real-Time Data Streaming:

- Enable continuous compliance checks by processing live data streams.
- Instantly flag violating transactions as they occur.

BI Tool Integration:

- Connect with Power BI, Tableau, or Looker for advanced data visualization.
- Create interactive dashboards for detailed analysis and reporting.

Thank You

PHHC