



Gen-AI Powered Data Profiling and Validation System

Automated Compliance Checking for Financial Data

Team: Hakuna Matata

Anshoo Rajput
Parva Patel
Neel Thakker
Mona Gupta

March 26, 2025

System Overview

Architectural Diagram:

Key Components:

Mistral-7B LLM

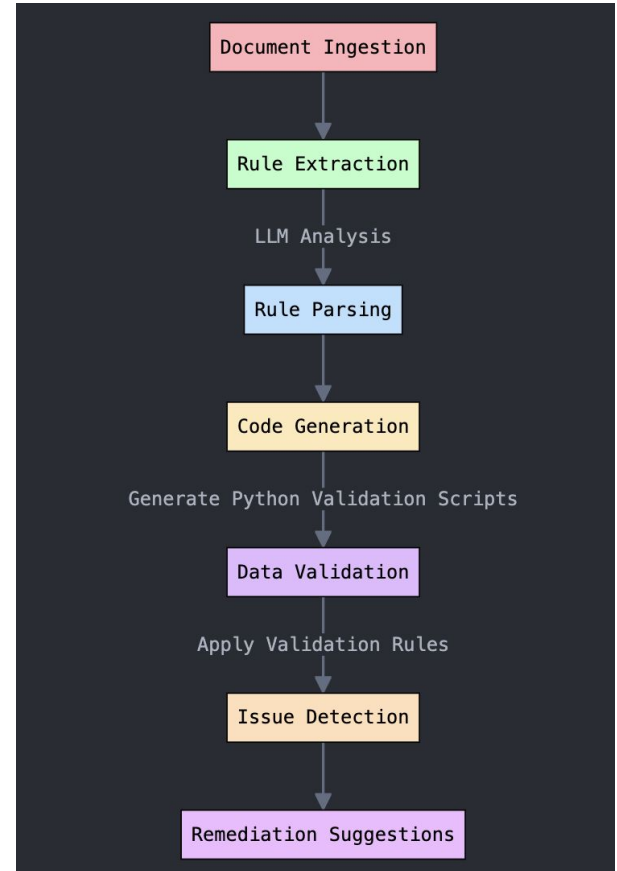
- Advanced AI model for extracting validation rules
- Generates precise Python validation code from regulatory documents

Vector Database

- Enables semantic search across compliance documents
- Supports efficient retrieval and indexing of regulatory information

SQLite Database

- Lightweight storage for validation rules and results
- Provides reliable local data persistence and querying





Model Selection - Mistral-7B

Performance Advantages

- Matches performance of 70B parameter models
- Highly efficient computational resource utilization
- Open-weight model with commercial usability
- Optimized for complex instruction following

Key Selection Criteria

- ✓ Cost-effectiveness
- ✓ Performance-to-size ratio
- ✓ Operational flexibility

Alternatives Comparison

- GPT-3.5/4: Superior but prohibitively expensive
- Llama 2: Larger versions resource-intensive
- Claude: Limited API accessibility



Technical Approach

1) Structured Rule Extraction

- Strict JSON output formatting
- Multi-level LLM response validation
- Ensures precise, consistent rule generation

2) Self-Correcting Code Generation

- Enforced function signature constraints
- Automated execution testing
- Guarantees code reliability and accuracy

3) Flexible Input Modes

- File upload (CSV/Excel)
- Manual data entry
- Supports diverse data input scenarios

4) Concurrency Management

- Thread-safe database operations
- Implements robust access control
- Prevents data conflicts in multi-user environments

AI Solution vs Traditional ETL Systems



Comparative Advantages

	Aspect	Traditional ETL	AI-Powered Solution
Key Impact 🚀 Reduces Rule Implementation Time: Days → Minutes	Rule Creation	Manual Coding	Automated Extraction
	Maintenance	High Complexity	Minimal Effort
	Adaptability	Weeks to Change	Minutes to Update
	Error Handling	Generic Messages	Contextual Remediation
	Cost Efficiency	Higher maintenance cost	Lower long-term cost



Implementation Challenges & Solutions

Technical Hurdles

- **LLM Output Consistency**
 - Multi-stage validation
 - Regex and schema checks
- **Code Generation Safety**
 - Sandboxed execution
 - Strict input/output contracts
- **Document Variability**
 - Intelligent chunking
 - Contextual overlap strategies

Operational Constraints

- Hugging Face rate limit management
- On-premise document processing
- Privacy preservation



Future Enhancements & Research

Product Roadmap

- Multi-document cross-referencing
- Rule versioning with temporal tracking
- Cloud data warehouse integration
- Domain-specific model fine-tuning

Research Directions

- Few-shot learning for rare rule patterns
- Automated test case generation
- Advanced contextual understanding






Conclusion: AI-Powered Compliance Revolution

Key Takeaways

- LLMs transform complex compliance workflows
- Structured prompting ensures reliable AI outputs
- Proven production-ready accuracy

Impact

-  From Manual to Intelligent Automation
-  Precision at Scale
-  Compliance Reimagined

Thank You!
