RegulAItion

An LLM-based model for banking regulation answers

Bridging Legal Text and Operational Decisions with Artificial Intelligence

The Challenge:
Regulatory Overload

- ***Reality* -** Bank clerks navigate complex yet mandatory regulations that shape their daily operations.

- ***Problem* -** Regulations are contained in documents that are hard to access and interpret, making their reliable application difficult.

- ***The Goal* -** Create a model that interprets rules and aligns them with available data for accurate regulation compliance.

Why It's Hard & Current Solutions

Core Challenges

- ***Fragmentation:*** Regulations scattered across multiple authorities.
- ***Complexity:*** Ambiguous definitions and dense legal jargon.
- ***Explainability:*** AI must tie every decision to a specific clause.

Existing Approaches

- ***Manual Work:*** Officers interpret docs and update guidelines by hand.
- ***Old Tools:*** Reliance on static spreadsheets and Word docs.
- ***Rigid Engines:*** Rule-based systems (e.g., Drools) require manual translation.

Project Vision & Novelty

The Solution

An LLM based model that delivers accurate, source-grounded answers to banking-regulation questions by retrieving and analyzing real regulatory documents.

Key Novelty

- Combines all rules into one place
- Understands the real meaning of the rules
- Gives clear, easy-to-explain answers.

System Specifications & Technology

Inputs & Outputs

Input:

- A free-form natural language query regarding banking regulatory questions.
- Database regarding regulatory documents.

Outputs:

- Yes or no answer.
- Grounded natural-language answer.
- Precise Source References (Context).
- Confidence Score.

Models & Techniques

- Document Parsing,
  Extracting clean text from PDFs and HTML files.
- Embedding Generation
  Turning text into numerical vectors the model can search and compare.
- Synthetic Q&A Creation
  Using a model to generate training examples from the documents.
- Data Validation
  Cleaning the dataset and splitting it into train/test sets.
- Fine-Tuning & Retrieval
  Training the model on the Q&A and using vector search to find relevant text for each question

Technical Pipeline

1

Collection

Ingest & Parse
PDF/HTML cleaning
(No ML)

2

Synthesis

Synthetic Q&A
using Teacher LLM
to create dataset

3

Validation

Data Cleaning
& Validation Split
(Train/Test)

4

Training

Fine-tune LLM
& Vector Search
Retriever

5

Evaluation

Metrics Protocol
& Baseline
Comparison

Data Specification & Generation

Synthetic Data Pipeline

- **Source:** Public regulatory texts (Basel, AML/KYC) split into chunks.
- **Generation:** A Large LLM acts as a "Teacher," reading chunks and generating Q&A pairs.
- **Labeling:** Every output includes the Question, the Answer, and precise Citation IDs.

Metrics & KPIs

Answer Accuracy

checks if the model's yes/no answer is correct; it works by matching the answer to the expected result.

Citation Accuracy

checks if the model used the right part of the regulation; it works by comparing the section the model cited to the correct one.

Confidence Score

shows how sure the model is, it works by measuring how similar the model's answer is to the retrieved text.