CAM DS 401 Employer Project

Final Report

Team: Byte & Sterling

Thomas Aaron

Jason Aspell

Joseph Church

Prasanth Pagolu

Nairio Santos

Giuseppe Tenaglia

Submission Date: 10th March 2025

Contents

Background and Problem Statement 3
Project Development Process 4
Workflow & Orchestration Agent 5
Data Processing and Storage 7
Credit Risk Scoring Methodology 8
Topic Modelling & Sentiment Analysis 9
RAG risk analysis and web search 12
Results and Roadmap 12
Appendix 14
Database Schema 14
Core Data Schema 14
Conversation Analysis Schema 15
NLP Processing Schema 16
Sentiment Analysis Schema 17
Search Function Schema 17

Background and Problem Statement

The Bank of England (BoE) regulates and supervises firms via its Prudential Regulation Authority, responsible for approximately 1,500 banks and other firms. Among these are Globally Systemically Important Banks (G-SIB), which are considered significantly important to the global financial system such that any failure would constitute a significant risk. The RegTech, Data and Innovation team analyses various data sources to provide useful insights, and is seeking to enhance its oversight of financial markets by identifying insights from quarterly results announcements and earnings call Q&A transcripts.

These data sets are particularly challenging to analyse, given the unstructured or inconsistent form of the data, or complex financial content. This project aims to address the limitations of traditional data science methods, incorporating NLP techniques such as sentiment analysis, topic modeling, and information summarization, tailored specifically for the financial sector. The techniques used will be evaluated and refined, providing alternate approaches to regulating these firms.

Insights will relate to the broader financial market as well as a smaller number of G-SIBs, ensuring the BoE can anticipate potential disruptions and mitigate the impact. The result will be unique, actionable insights beyond traditional financial risk models, through accessible and interpretable AI models.

Project Development Process

Following agreement of the core objectives, determining the approach was critical to shaping the process. This was undertaken by considering existing experience within the team whilst researching new, unfamiliar approaches, weighted against the potential opportunities and risk for the project.

Preliminary ideas were stored in an ideation matrix, with weight given to impact, cost, effort, time, and quality of each technique, alongside the contribution to each risk type (detailed later).

This allowed various strategies to be evaluated, and the problem approached in the optimum way.

Once scores were allocated, modelling commenced with those scoring the highest. The matrix is displayed at Fig. 1.

Idea	Priority
Evaluate other methods besides Spacy	1
FinGPT to analyse sentiment	2
RAG+Openai	3
FinBERT	4
Generate own training data using RAG	5
FinBert for sentiment analysis	6
LLM for sentimental analysis	7
BertTOPIC	8
Topic modelling with Bertopic	9
Seeking Alpha to download the transcripts	10
FinGPT - Data failed banks	11
FinGPT - A/B testing	12
Collect and compare Interest Rates	13
Input data in Gemini and query answers	14
Use openAI to read the transcripts and query for topics and sentiment	15
LLM to structure data	16
Analyse Tariffs impact	17
FinGPT - Regulatory analysis	18
Get data back to 2008	19
LDA/NMF for topic modelling	20

Figure 1: Ideation Matrix

Workflow & Orchestration Agent

In addition to prioritising potential techniques, a clear workflow was necessary to ensure a coherent start-to-end process. Following the processing of transcripts with NLP and LLM models, establishing advanced storage methods would ensure efficiency and quality during downstream analysis. The workflow structure is shown at Fig. 2.

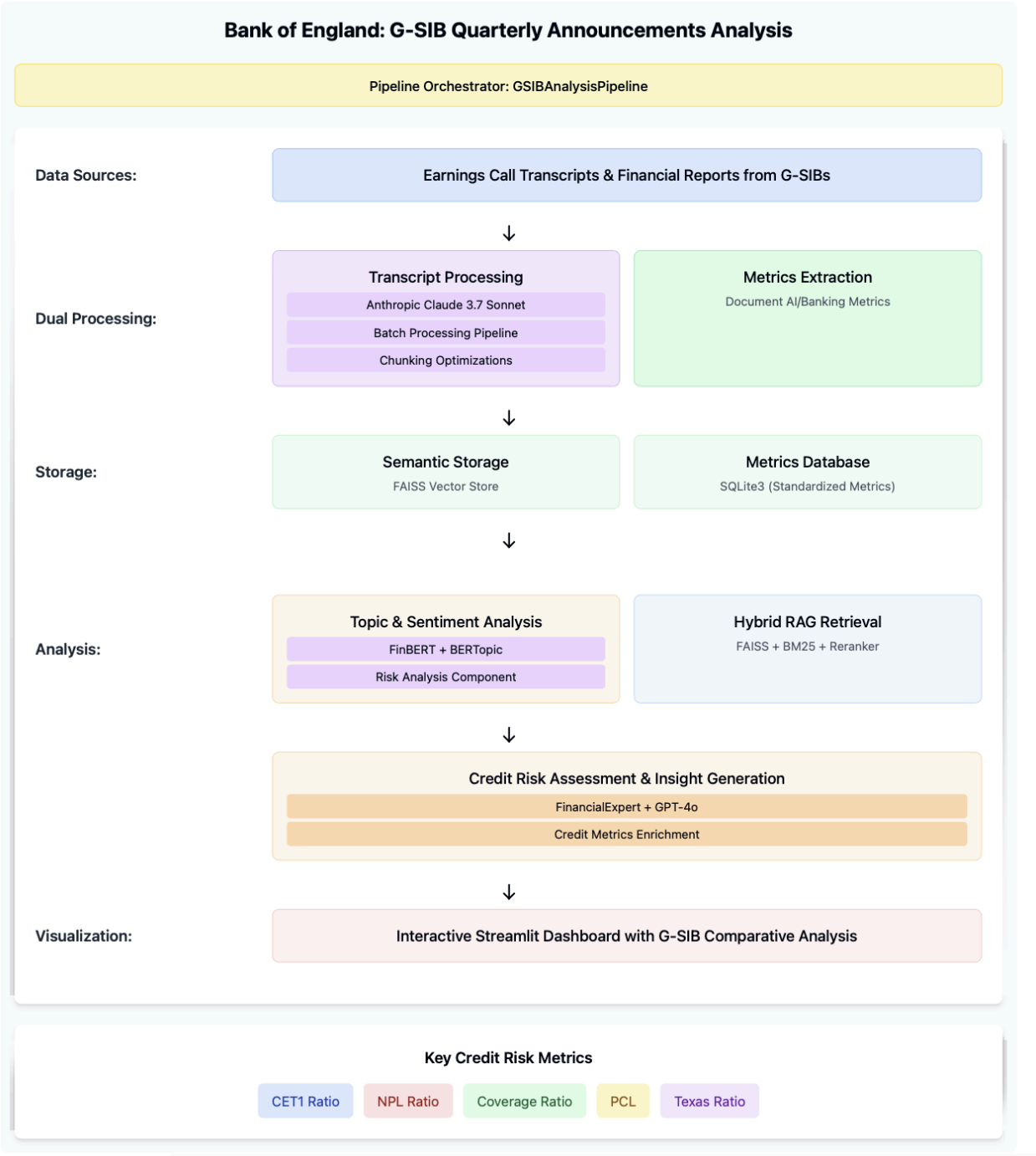


Figure 2: Solution Structure

Analysis was undertaken on earnings reports, analyst calls, and Q&As, with advanced orchestration utilised to achieve multiple data processing tasks. The orchestration agent is responsible for managing the end-to-end pipeline of the transcript processing as follows:

- Automated Data Retrieval: downloading the transcripts and financial reports.
- Text Processing and NLP: extracting key financial objects using NER and cleaning transcript data through LLM.
- Database Storage and Retrieval: storing questions and answers with the necessary details in an efficient database for analysis.
- Parallel Execution and Optimisation: processing transcripts simultaneously, improving efficiency.

The orchestration agent is an essential element of the workflow, automating data processing and managing the analysis method, improving accuracy and efficiency, and allowing large volumes of data to be analysed.

Data Processing and Storage

To convert unstructured earnings call transcripts into structured data, our system now utilizes the Anthropic Claude API for comprehensive NLP processing. Large Language Models (Sonnet 3.7) handle the entire extraction pipeline—from speaker identification to Q&A pair matching to text cleaning—within a unified framework. The system processes transcripts through intelligent chunking with overlap, enabling context-aware analysis even with lengthy documents. Batch API processing facilitates efficient handling of multiple transcript segments simultaneously, while database integration ensures persistent storage with robust transaction management. Advanced entity recognition, contextual understanding, and topic identification capabilities accurately capture financial discussions and key metrics, delivering high-quality structured data for downstream analysis.

Meanwhile, metric extraction focuses on standard credit risk metrics from G-SIB reports, including NPL, Coverage, PCL, CET1, and Texas ratio. The custom-trained Document AI solution achieves upwards of 95% accuracy, compared to only 50% with traditional NLP methods for similar metrics. This is alongside the ability to process multiple reports simultaneously, with minimal maintenance when formatting of input documents changes.

The storage layer combines relational database storage with FAISS vector embedding to achieve comprehensive text analysis. The SQLite database provides a structured management of transcripts, metrics, topics, and sentiment, through an established schema (Fig. 3). The FAISS vector storage complements this, with HNSW indexing enabling powerful semantic searching with cross-encoder reranking. This integrated approach produces a robust foundation for both precise data management and advanced semantic retrieval capabilities in financial text analysis.

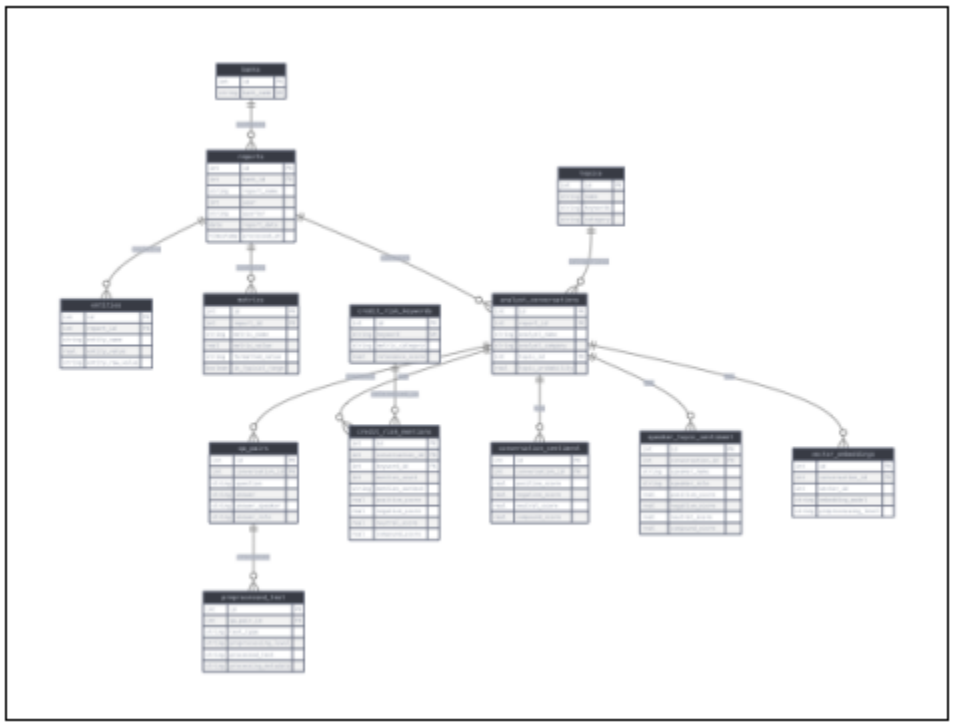


Figure 3: Database Schema (see Appendix for detail)

Credit Risk Scoring Methodology

Extracting financial metrics from lengthy G-SIB reports initially achieved only 50% accuracy using traditional NLP methods (Camelot, NER, transformers), requiring frequent maintenance as report formats changed. The present project employs custom-trained Google Document AI that first identifies relevant pages containing metrics, significantly improving processing efficiency. This approach achieves 95%+ accuracy through precise manual labeling and targeted training data, providing the structured metrics foundation for the credit risk scoring system without

requiring extensive code changes when report formats evolve. The ‘Credit Risk Analyser’ class evaluates the credit risk of G-SIBs using a multi-factor methodology, combining financial metrics, sentiment analysis, and risk adjustments:

1. **Financial metrics:** Five key financial indicators from quarterly reports are analysed: Non-Performing Loan (NPL) ratio, Coverage Ratio, Provision for Credit Losses (PCL), CET1 Capital Ratio, and Texas Ratio. Increased NPL, PCL, and Texas Ratio values suggest higher risk, while lower Coverage and CET1 Capital Ratio value lower risk. Each metric is scored on a scale of 10-100.
2. **Sentiment analysis:** NLP is used to gauge sentiment, with negative sentiment suggesting an increased risk level and a positive sentiment a lower risk level.
3. **Risk adjustments:** Adjustments for size and volatility are considered. With larger banks weighted more heavily on NPL and PCL, recent volatility in financial metrics amplifies the magnitude of risk.

The final Credit Risk Score is a value from 0-100, categorized on a scale from ‘Very Low’ to ‘Very High’ (Fig. 4).

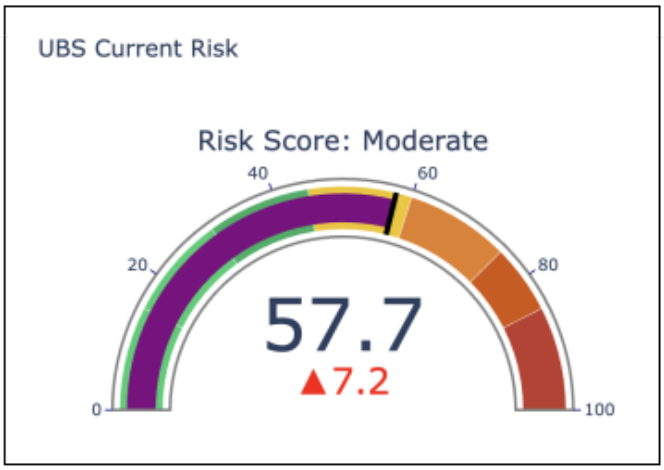


Figure 4: Overall sentiment score

Topic Modelling & Sentiment Analysis

Following text cleaning via the LLM, the analysis pipeline processes earnings call transcripts through several specialised stages. The preprocessing phase employs a custom financial

domain pipeline preserving critical terms (NPL, CET1, provisions) while expanding abbreviations and performing entity-aware lemmatization. This domain-specific approach ensures financial terminology integrity throughout the analysis.

The topic modelling implementation leverages BERTopic with financial-optimized parameters (UMAP n_neighbors=3, min_dist=0.0; HDBSCAN min_cluster_size=2) to identify earnings call themes. Conversations are embedded and clustered before being mapped to risk categories through vector similarity comparison with predefined financial risk embeddings, assigning risk relevance scores to each cluster.

Sentiment analysis employs the FinBERT-tone model to analyse financial sentiment at multiple levels: conversation-wide, individual QA pairs, and speaker-specific assessments (Fig. 5). The architecture captures sentiment hierarchically, with particular attention to executive responses in risk-related discussions.

Figure 5: Sentiment score by speaker

Risk-sentiment correlation is achieved by integrating topic categories with sentiment scores (Fig. 6), enabling multi-dimensional analysis across risk categories, financial institutions, and

reporting periods (Fig. 7). This approach reveals sentiment patterns and outliers within specific risk domains.

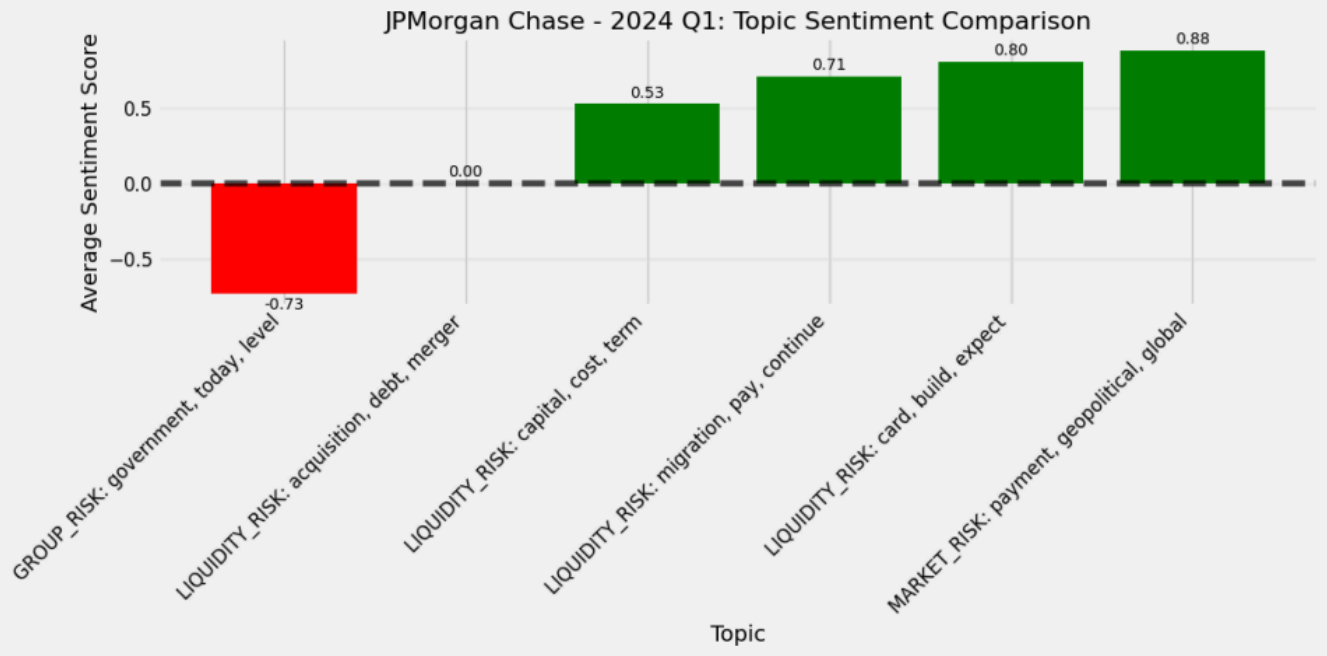


Figure 6: Sentiment score by topic

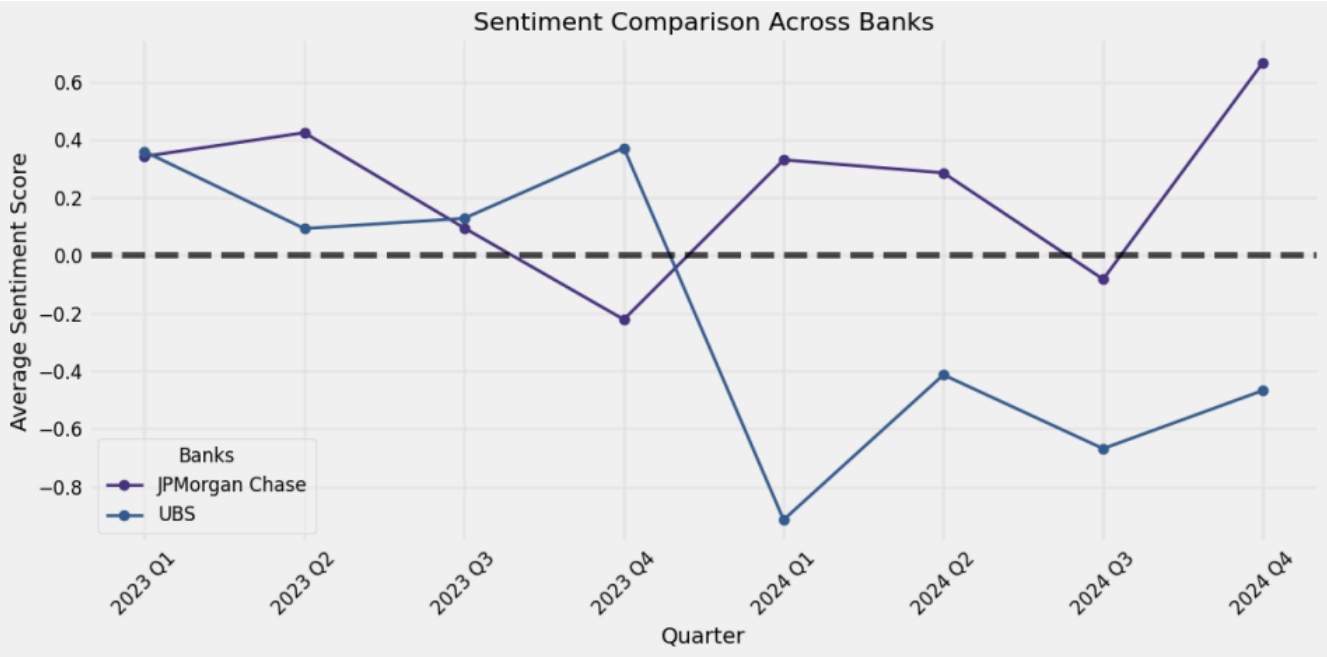


Figure 7: Sentiment score by bank over time

These methods come with inherent challenges when targeting financial language processing, particularly the nuanced terminology around credit metrics and provisions that standard NLP models may misinterpret without domain adaptation. The custom preprocessing and financial-specific embeddings incorporated into this model mitigate these limitations, but cannot entirely eliminate them.

RAG risk analysis and web search

The RAG model is designed to summarise key insights from a transcript when given a query containing a specific risk type, financial quarter, and bank. FinLANG embedding is used for financial embeddings, while GPT-4o acts as a generative model. Retrieval uses FAISS HNSW as dense retrieval and BM25 as sparse retrieval, in which specific keywords are selected based on the type of risk to be analysed. These retrieval methods are combined via reciprocal rank fusion, with MS MARCO as the reranker.

After the query is passed to the model and retrieval complete, the top 10 contexts are summarised with GPT-4o to extract key insights.

The sample pipeline uses RAG to extract insights from JPMorgan and UBS, relating to credit risk from quarters 2, 3, and 4 in 2024. A full report is provided on Q4 2024, alongside an evolution of credit risk factors over this time period. GPT-4o then provides a quantitative evaluation (from very low to very high), providing a snapshot of the level of risk the bank is facing.

To effectively validate the output of the RAG model, a web search is run on historic quarterly financial news reports. Google Cloud search and Server.dev are employed as search engines, isolating financial news in Reuters, Financial Times, and Bloomberg. This data is often unavailable without paid access, so Bright Data API is utilised to obtain summaries of relevant articles. Finally, GPT-4o summarises the RAG insights, and returns an alignment score determining how closely the content of the articles align with the output of the model.

Results and Roadmap

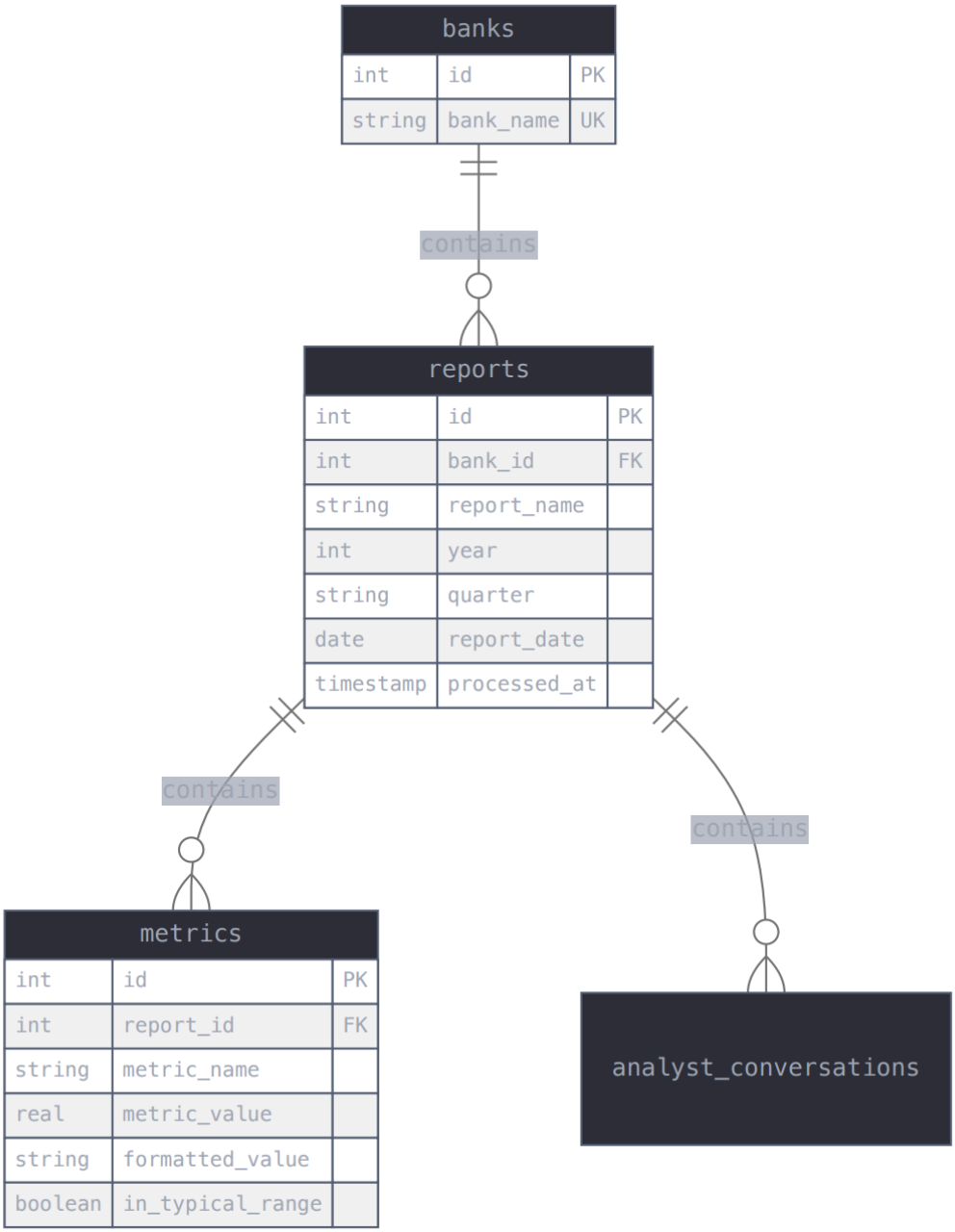
With ByteSight the team has demonstrated a comprehensive sentiment and insight workflow, successfully completing a proof of concept with UBS and JPM. The analysis shows UBS's sentiment score dropping significantly from Q4 2023 to Q1 2024, driven by rising credit and group risk, while JPM remains stable. UBS also exhibits a consistently higher Risk Score.

The roadmap for H1 2025 will focus on integrating the backend data generation to front end visualisation in the Streamlit.app, alongside evaluating additional risk factors and extending G-SIB coverage. By H2 2025, additional data sources and chatbot integration are planned. Longer-term goals for 2026 include expanding to other sectors, ESG integration, live quarterly analysis, and regional and video-based insights.

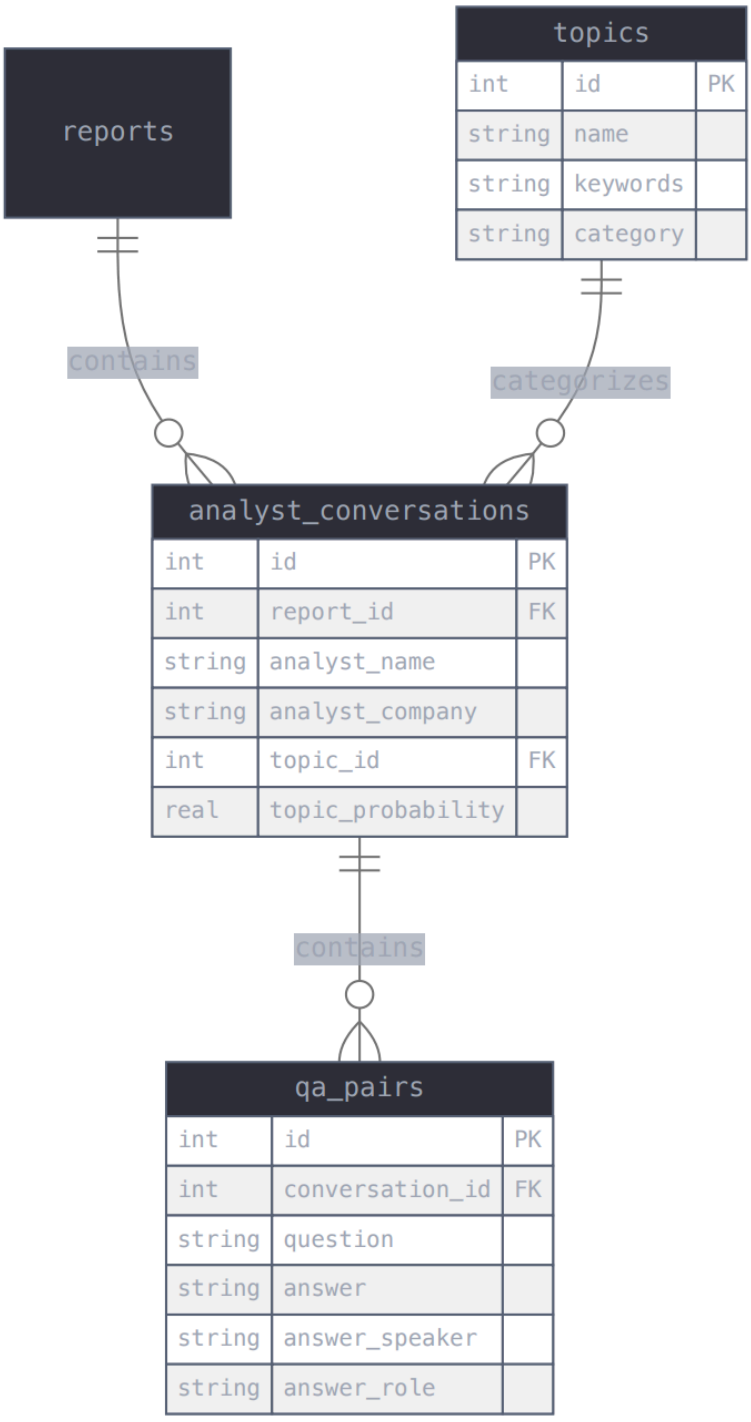
Appendix

Database Schema

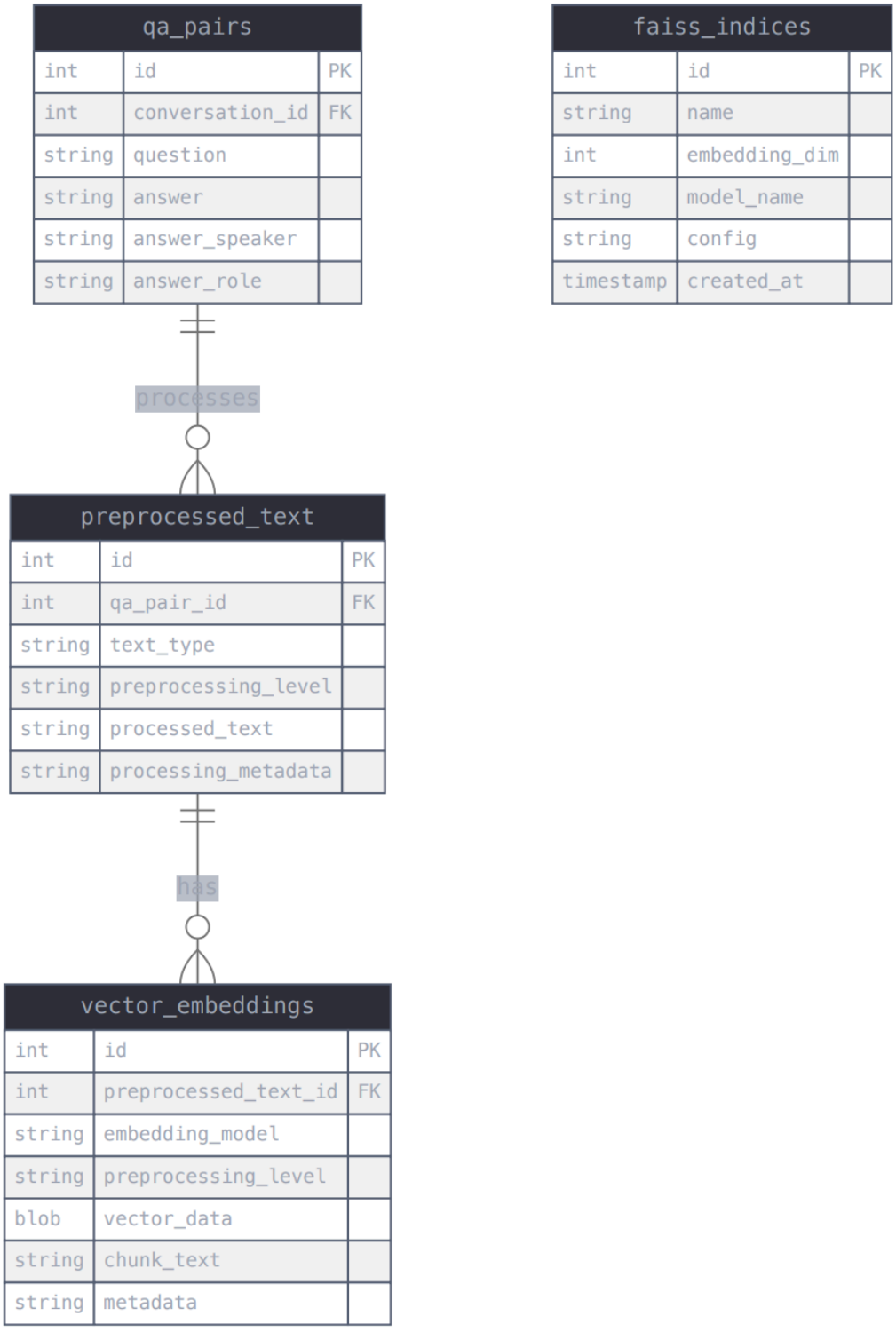
Core Data Schema



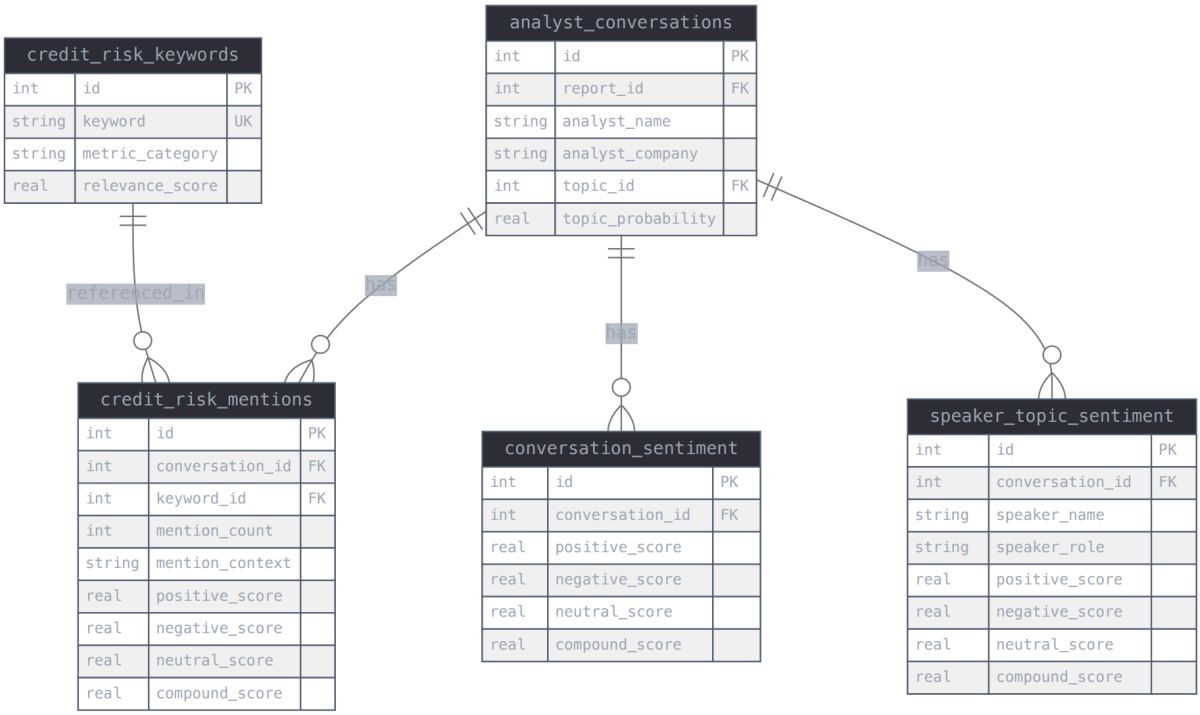
Conversation Analysis Schema



NLP Processing Schema



Sentiment Analysis Schema



Search Function Schema