From Transcripts to Trends: Applying NLP to Bank Q&As

Employer Project: Bank of England

David Revell
Fabio Mufañas
Luc van den Brink
Marta Santana Silva
Nicolas Voltier
Simba Prew

Introduction

The Bank of England, through the Prudential Regulation Authority (PRA), oversees more than 1,500 financial institutions to safeguard financial stability. A vital aspect of this role is analysing quarterly earnings announcements and Q&A sessions, where executives and analysts discuss results, risks, and strategy. These transcripts contain valuable signals but are long, unstructured, and resource-intensive to interpret, making it difficult to systematically track thematic changes or shifts in tone.

This project applies natural language processing (NLP) to tackle that challenge. Using Barclays and Citibank transcripts, we developed a pipeline that extracts structured Q&A pairs, classifies themes, measures sentiment and orientation, and scores executive responsiveness. These outputs provide a scalable view of how discussions evolve across time and firms. Complementing this, a retrieval–augmented generation (RAG) chatbot ingests all transcripts, allowing supervisors to query narratives interactively with prompts informed by the classical pipeline. Together, these tools enhance supervisory monitoring, benchmarking, and early risk detection.

Methodology

CLASSICAL APPROACH

1 - Data ingestion and structuring

The first task was to transform raw PDF transcripts into structured Q&A data. Barclays and Citi documents vary in formatting, containing headers, disclaimers, and inconsistent speaker labels, so a robust parsing process was required. We used PyMuPDF, which provided stable extraction and fallback options when text layouts differed across years. Metadata such as bank, year, and quarter were parsed from filenames to enable chronological analysis.

A major challenge was speaker detection. We built global rosters of executives and analysts by scanning all transcripts, then used these rosters to validate potential speaker lines. Dialogue was segmented into turns, aggregating text until the next recognized speaker.

From there, we implemented a turn extractor that aggregated all lines following a speaker label until the next label. For Barclays, pairing was straightforward: analyst turns (questions) were paired with subsequent executive turns (answers). For Citi, the Q&A boundary often lacked explicit markers, so we defined the Q&A section either at the explicit “Q&A” heading or, failing that, the first non-executive/host speaker.

Finally, we applied a pairing function: each analyst's question was paired with consecutive executive answers, with operator/host turns discarded.

2 - Single pair extraction and Analytical Features

From the raw Q&A blocks, we extracted single structured question–answer pairs and annotated them with additional analytical attributes:

- **Degree of answering** – a measure of how directly an executive responded to the analyst's question.
- **Orientation** – whether the question referred to past results, forward-looking guidance, or both.
- **Standardized theme** – classification of question content into regulatory-relevant categories (e.g., risk, capital, profitability).
- **Emotion and sentiment** – fine-grained tone labels at both question and answer level.
- **Topics** – unsupervised clusters of thematically similar content.

2.1 - LLM-based extraction

The best way to extract individual questions and answers turned out to be by using an LLM. Initially, Phi-4-mini-instruct was used with few-shot prompting since it could be run locally. However, the results from that model were insufficient and still required significant computational resources.
Eventually, the GPT-5-mini and GPT-5-nano models from OpenAI were used, which is possible since the Q&A transcripts are public data. Since the question and answer extraction task is quite complex, the GPT-5-mini model was used, which has a balanced trade off between size, speed and cost. To save costs, with a single fine-tuned prompt, the questions and answers were extracted, and the degree of answering of the question was determined. Combining this into a single prompt compared to using separate prompts did not result in significant quality differences for this model.
To determine the question orientation and standardized theme, the smaller, faster and cheaper GPT-5-nano turned out to be sufficient and consistent with a single prompt. Overall the costs were minimal and computation time was manageable, especially considering quarterly results transcripts would only need to be processed four times a year. The theme determination is especially troublesome for the Citi group Q&A data because of more questions using context from previous questions or answers, compared to Barclays where an analyst asks one or multiple questions, gets an answer and mostly does not follow up.

2.2 - Topic modelling (FinBERT)

We applied advanced topic modelling to financial Q&A data, aiming to capture and compare underlying themes in questions and answers. Semantic representations are built using FinBERT ("ProsusAI/finbert"), a transformer pre-trained on financial text, integrated into a SentenceTransformer with pooling to generate robust embeddings.

Data preparation includes joining extracted nouns into coherent text segments and applying regex-based cleaning to remove filler expressions (e.g., “give us”, “sort of”). In addition, a comprehensive stopword set is constructed by combining standard NLTK stopwords with conversational and finance-specific terms, ensuring that thematic modelling focuses on substantive content rather than generic vocabulary.

Text vectorisation is performed through a CountVectorizer configured for unigrams and bigrams, with adaptive thresholds for document frequency and a cap of 18,000 features. To reduce dimensionality while preserving semantic relationships, UMAP projects the embeddings into a five-dimensional space using cosine distance.

Clustering is then conducted with HDBSCAN, which identifies dense, hierarchically consistent clusters without requiring a predefined number of topics. Finally, BERTopic integrates FinBERT embeddings, the refined vectorizer, UMAP, and HDBSCAN to generate interpretable topic representations.

2.3 - BERTopic integration

BERTopic combined these steps into interpretable topics with representative keywords. We ran it separately on questions and answers, enabling comparison of what analysts focused on versus what executives emphasised. Finally, topics were mapped back to standardised themes using a zero-shot classifier, ensuring alignment with the curated regulatory categories.

2.4 - Sentiment and emotion analysis

This sentiment analysis approach leverages a transformer-based pipeline to classify emotions within financial Q&A data. Specifically, the model *“bhadresh-savani/bert-base-uncased-emotion”* is employed, returning probability distributions across multiple emotion categories. Each text is processed in batches, ensuring computational efficiency, and the most likely emotion per entry is selected along with its confidence score.

To retain interpretability, the full probability distribution of emotions is also stored, enabling both categorical and probabilistic analyses. The system applies classification separately to questions and answers, thus allowing for a direct comparison of emotional tone across the two dimensions of dialogue.

Beyond raw emotion classification, the method incorporates a mapping function that translates emotions into broader sentiment categories: positive (*joy, love, surprise*), negative (*anger, sadness, fear, disgust*), neutral, or uncertain when scores fall below a defined threshold. This step provides a higher-level perspective aligned with typical sentiment analysis practices.

By combining fine-grained emotional detection with aggregated sentiment mapping, the framework offers a nuanced view of how financial stakeholders express themselves in Q&A interactions

RETRIEVAL-AUGMENTED GENERATION (RAG)

The aim was to build an agent capable of retrieving context from pre-ingested documentation, including Barclays and HSBC transcripts and PRA rulebooks.

1 - Data preparation for RAG

Documents were bundled into a corpus (e.g. PRA or individual banks), with each document and page summarised. Key facts were extracted per page, and paired with generated questions to create retrievable QA links. Embeddings were generated for both facts and questions to enable cosine-similarity retrieval. The resulting structure — including facts, pages, and their links — was persisted as a graph in Neo4j (see Figure 1).

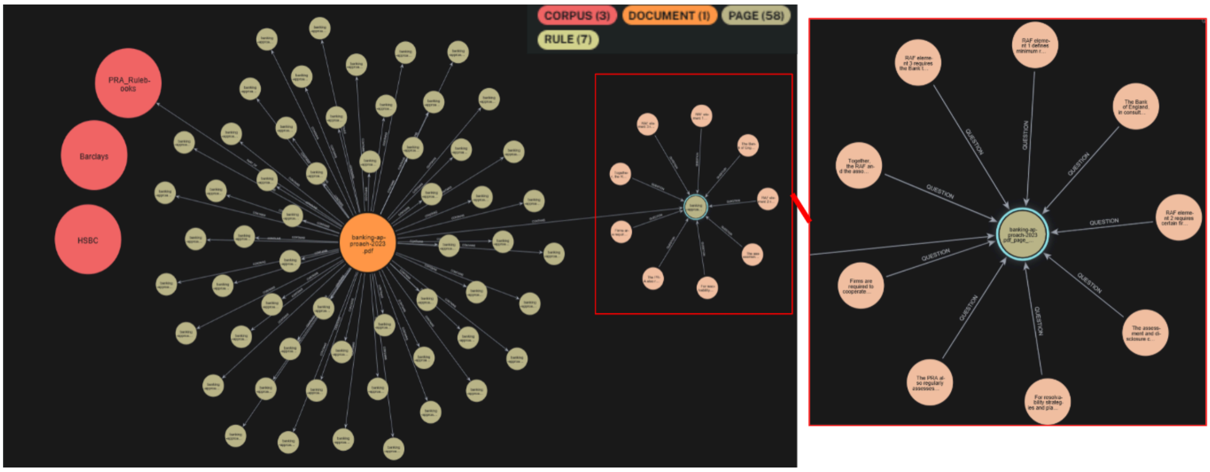

Figure1: Neo4j grah: summaries and embeddings are stored as node properties.

2 - Contextualized Agent

The agent used three tools: `documentation_overview_tool` (for general summaries), `RAG_tool_PRA` (to retrieve relevant rules), and `RAG_tool_Banks` (to fetch bank-specific facts). The conversation flow prioritised querying PRA rules to clarify the user's question, followed by searching the bank documentation. As part of retrieval, references were collected and attached automatically to the agent's responses, supporting traceability.

EMPLOYEE:
Hi, i want to investigate capital and liquidity

CHATBOT:
Here are PRA-informed enhancements to frame your capital and liquidity inquiry, ensuring we cover core PRA expectations:

- Capital framework and planning
  - Distinguish Pillar 1 minima, Pillar 2A, the combined capital buffer, and the PRA buffer. (Four-part capital framework; internal assessment still required by the firm.)
  - Emphasise ICAAP alignment, prudent capital assessment, and not reliance on regulatory minima or aggressive accounting. (Pillar 1/2A, buffers, ICAAP guidance)
  - Ensure capital quality aligns with CRR/SS7/13 criteria for capital instruments.

- Liquidity framework
  - Require a framework that captures the full range of liquidity risks, with stress testing of those risks; PRA will assess whether the stresses are prudent. (Liquidity framework and stress testing)
  - Cover LCR and NSFR considerations and contingency funding planning as part of liquidity management.

- Governance and risk management
  - Board and senior management design and effectiveness; robust risk controls and governance around capital and liquidity decisions.
  - Explicit articulation of how risk management informs capital adequacy assessments. (Governance and controls)

- Supervisory context and reporting
  - PRA’s forward-looking assessment approach and how capital/ liquidity decisions feed into supervisory planning and potential adjustments.

If you want, I can now pull specific bank documents (HSBC/Barclays) to map how these PRA expectations are implemented in practice.

Next steps: tell me which banks or documents to prioritise, or confirm you want a general PRA-to-bank mapping first.

PRA References:
  o PRA_Rulebooks - banking-approach-2023.pdf: page n°021, page n°028, page n°029, page n°003, page n°054, page n°046, page n°027, page n°009, page n°034, page n°002

CHATBOT:
Here is HSBC-specific PRA-informed mapping to capital and liquidity expectations, with direct evidence from HSBC materials and PRA rulebooks.

1) Capital framework (Pillar 1 minima, Pillar 2A, combined PRA buffer, PRA buffer; ICAAP alignment; capital quality)
- HSBC evidence: Pillar 2A requirement as of 31 Mar 2025 was 2.6% of RWAs, with 1.5% required to be CET1; HSBC compliant through Q1 2025. This aligns with PRA’s IFR/ICAAP framework and PRA’s emphasis on Pillar 2A as part of ICAAP alignment. (Bank evidence: HSBC earnings release, Pillar 2A 2.6%; page 046) [Bank fact: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:894]
- PRA framework references (for planning and structure): Pillar 1, Pillar 2A, the combined capital buffer, and the PRA buffer form the four-part framework; PRA buffer may apply in addition to minima; capital instruments must meet CRR/SS7/SS13 criteria. (PRA rulebook context) [PRA facts: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:3139, 3150, 3155]

2) Liquidity framework (full liquidity risk framework, stress testing; LCR, NSFR; contingency funding planning)
- HSBC evidence: LCR was 139% at 31 March 2025 (HSBC notes LCR exposure; a related factual line shows LCR values). This evidences a formal liquidity framework including LCR oversight. (Bank fact: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:889, 980)
- PRA framework references (LCR/NSFR/stress tests; contingency funding): PRA emphasizes a framework that captures full liquidity risks with stress testing that is prudent, and coverage of LCR, NSFR, and contingency planning. (PRA rulebook context) [PRA facts: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:3063, 2892, 2894, 3445]
 release. (Bank facts: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:950, 970)
- PRA framework references: Clear emphasis on Board and Senior Management design/effectiveness; risk controls; and explicit articulation of how risk management informs capital adequacy assessments. (PRA facts: 4:e43eb764-a2c6-44ce-993e-d18abbf24318:2894, 2893, 3023)

4) Supervisory context and reporting (forward-looking assessment; how capital/liquidity decisions feed supervisory planning)
- HSBC/Bank evidence: Forward-looking risk and capital considerations are embedded in the reporting and commentary around capital adequacy; credit flows and liquidity stress context. (Bank facts: 71, 894, 950)
- PRA framework references: PRA’s forward-looking assessment approach to determine capital needs and how decisions feed supervisory planning; integration into supervisory processes. (PRA facts: 3063, 3445)

5) International supervision context (home/host supervision; SS5/21; equivalence considerations)
- PRA framework references: PRA guidance on branch/subsidiary supervision and international banks emphasizes SS5/21 and related home/host supervision and equivalence considerations. (PRA facts: 3413, 3414)

6) Data and evidence requirements (quantitative and qualitative data; governance documents; risk appetite statements; ICAAP reports)
- PRA framing notes emphasize gathering a broad data set for ICAAP and capital planning in a forward-looking, risk-based way. (PRA facts: 3063, 3150)

If you want, I can now pull HSBC UK-specific documents to map these PRA expectations to practice in more depth (prioritising HSBC UK operations or a broader PRA-to-bank mapping). Refine or check anything else?

BANK References:
  o HSBC - 250429-1q-2025-earnings-release.pdf: page n°051, page n°046

3 - Retrieval Evaluation

In the final evaluation stage, we assessed retriever performance using RAGAS, which compares retrieved contexts against references to compute precision and recall. Queries ran against a Neo4j graph of atomic facts linked to their page and source document. Results were ranked by cosine similarity and filtered with a threshold (τ), using a per-document top-k = 5 to preserve provenance and avoid flooding. We varied τ and standardised on ≈ 0.7 as a balance between coverage and noise. Omission tests confirmed irrelevant facts weren't returned. All parameters and code are documented in the evaluation notebook and GitHub repo.

USER INTERFACE

1 - StreamLit

In order for the banking supervisors to be able to interact with the data and RAG, a Streamlit dashboard was created with two pages. On the first page, the supervisor is able to select various types of visualisations, filters, and groupings in order to get insights from the questions and answers. It is for example possible to easily check on which topics questions are being avoided, or which topics get more negative questions asked over time for example. On the second page, the supervisor can interact with the RAG, possibly asking questions about insights gained from the visualisations.

Results

1 - Classical NLP approach

Once the transcripts were transformed into structured and analyzable data, the classical NLP pipeline successfully enriched each question-answer pair with multiple analytical layers, guiding banking supervisors toward key areas of interest that may require closer examination.
Topic modelling using FinBERT embeddings and BERTopic revealed clear thematic shifts over time, such as increased analyst focus on profitability and risk management in 2021-2022 for Barclays, and more forward-looking discussions around capital and liquidity in later quarters. Sentiment analysis indicated that negative tones clustered around risk-related topics, while positive sentiment aligned with guidance and growth discussions (Figure2).

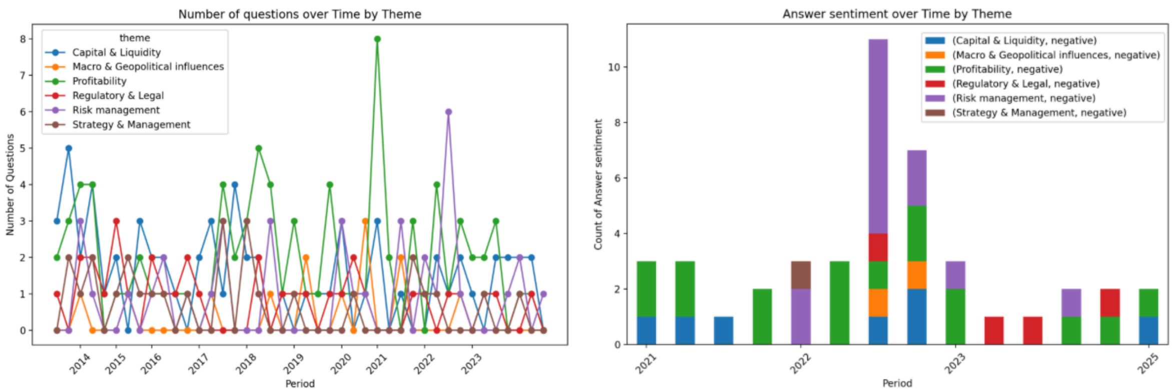

Figure2: Question and answer trends over time and theme (StreamLit).

These outcomes demonstrate that applying NLP to financial transcripts can uncover evolving areas of concern and interest without manual review.

However, limitations remain: theme and avoidance accuracy could improve by adding conversation-level context, fine-tune sentiment models on finance-specific data, and automatic detection of “areas of interest” in the dashboard would reduce manual exploration. The next step is integrating these insights more tightly into the RAG agent using derived topics and sentiments to refine retrieval prompts and support more targeted supervisory questioning.

2 - RAG

This evaluation tested the retriever’s accuracy in isolation. With a threshold of $\tau \approx 0.7$ and top-k = 5, the 20-question Barclays run yielded recall around 0.5, with slightly lower precision. Some questions returned no relevant context, while others succeeded. Higher thresholds improved precision but reduced coverage; lower ones added noise. An omission test set returned no contradictions but exposed limits of recall/precision when the correct answer is “not mentioned.” Runtime caps avoided unstable runs but didn’t affect results. One improvement is to restrict retrieval scope, mimicking how a chatbot might clarify context.