# Uncovering Risk in Bank Earnings Calls

## Using NLP to detect **early warning indicators of risk** in Q&A transcripts

## FinSight Project Report
## October 2025

**Team FinSight**

Lauren Brixey
Jerome Ahye
Anastasia Krivenkovskaya
Heidi Santos
Jyoti Rathod
Parush Parushev

UNIVERSITY OF
CAMBRIDGE

# 1. Background

The Prudential Regulation Authority (PRA) monitors financial stability through quarterly disclosures and earnings call transcripts from GSIBs. While structured reports capture key figures, they often miss tone, sentiment, and implicit concerns. Analyst Q&A exchanges can reveal evasiveness or emerging risks not evident in the data. However, the PRA currently lacks tools to systematically monitor these transcripts at scale. This project investigates whether natural language processing (NLP) can detect early warning signals in Q&A transcripts, enabling the PRA to prioritise higher-risk disclosures and deliver faster, data-driven supervisory insight.

# 2. Project Development Process

## 2.1. Problem Framing & Initial Exploration

The key challenge was to detect and quantify signals of risk, including tone, sentiment and evasiveness embedded in banks' Q&A transcripts. Verbatim exchanges from quarterly earnings calls were extracted and converted from PDFs into structured datasets, enabling systematic analysis and mapping to supervisory risk categories.

The analysis of these transcripts focussed on four core components:

- **Topic Modelling:** What themes dominate Q&A discussions, and how do they align with supervisory risks?
- **Sentiment Analysis:** How does sentiment/ tone vary across responses?
- **Evasion Detection:** Can instances of banker evasiveness be detected and how are these distributed across risk categories?
- **Summarisation:** Can outputs be summarised into concise PRA-aligned insights to support supervisory review at scale?

Given these analytical demands, we developed a structured NLP pipeline to move from raw transcripts to actionable outputs (**Figure 1**).

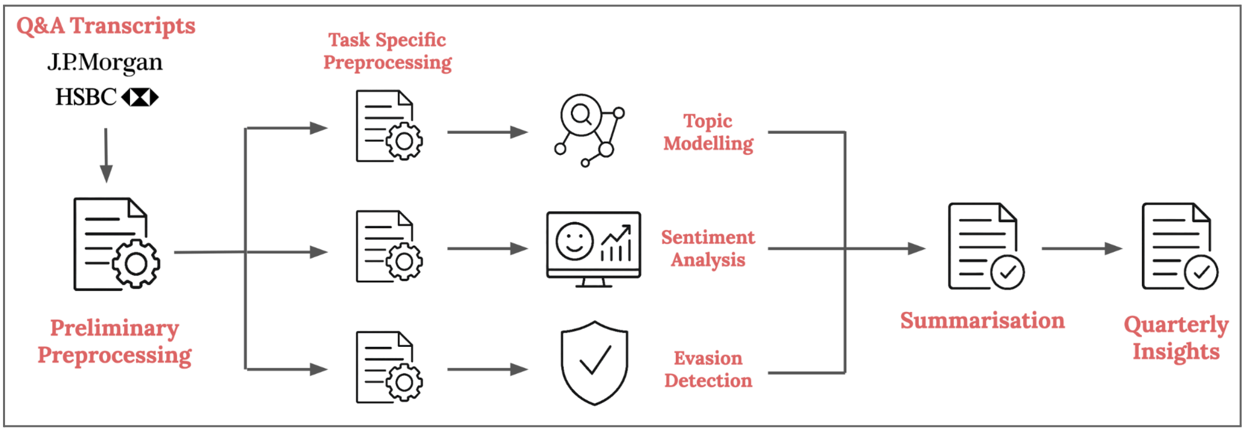


**Figure 1:** Diagram of the NLP Pipeline.

## 2.2. Data Preparation & Preprocessing

Transcript PDFs from JP. Morgan and HSBC were parsed into structured datasets and cleaned to remove headers, footers and filler text while preserving Q&A structure. Each record was standardised with speaker, role, firm and quarter using custom dictionaries for consistency, and pleasantries, duplicates and anomalies were filtered. This produced a reliable foundation for topic, sentiment and evasion analysis.

## 2.3. Topic Modelling

We explored both traditional and embedding based topic modelling approaches, including LDA, Top2Vec and BERTopic, to determine the most effective method for unstructured financial text. Classical models such as LDA and Top2Vec generated coherent topics but lacked flexibility for domain specific language, requiring extensive preprocessing to achieve financial context understanding.

Benchmarking identified BERTopic as the most suitable model due to its configurable parameters and support for embedding-based customisation. Integrating Finance2 transformer embeddings (vs FinBERT) improved semantic comprehension, while guided modelling using PRA-aligned seed keywords anchored clusters around regulatory risk categories (Ogunleye et al., 2023). Some categories, however, remained underrepresented due to sparse language in the transcripts.

A key challenge was managing filler and repetitive language, which distorted topic coherence and hindered PRA mapping. This was mitigated through custom stopword lists, guided clustering and embedding-based similarity matching, which refined topic quality and interpretability (**Figure 2a/b**). Nevertheless, manual review remained essential with this semi-supervised framework to ensure coherent outputs for downstream sentiment and summarisation tasks.

Our results demonstrate that topics discussed in quarterly calls can be modelled and grouped into regulatory relevant subjects. Topics categorisation was not straightforward due to the nuanced nature of the Q&As presented by the analysts and bankers. To manage this, we mapped each topic to the top three most probable PRA categories (**Figure 3a/b**), enabling the creation of interpretable and mappable topic clusters (**Figure 4**).

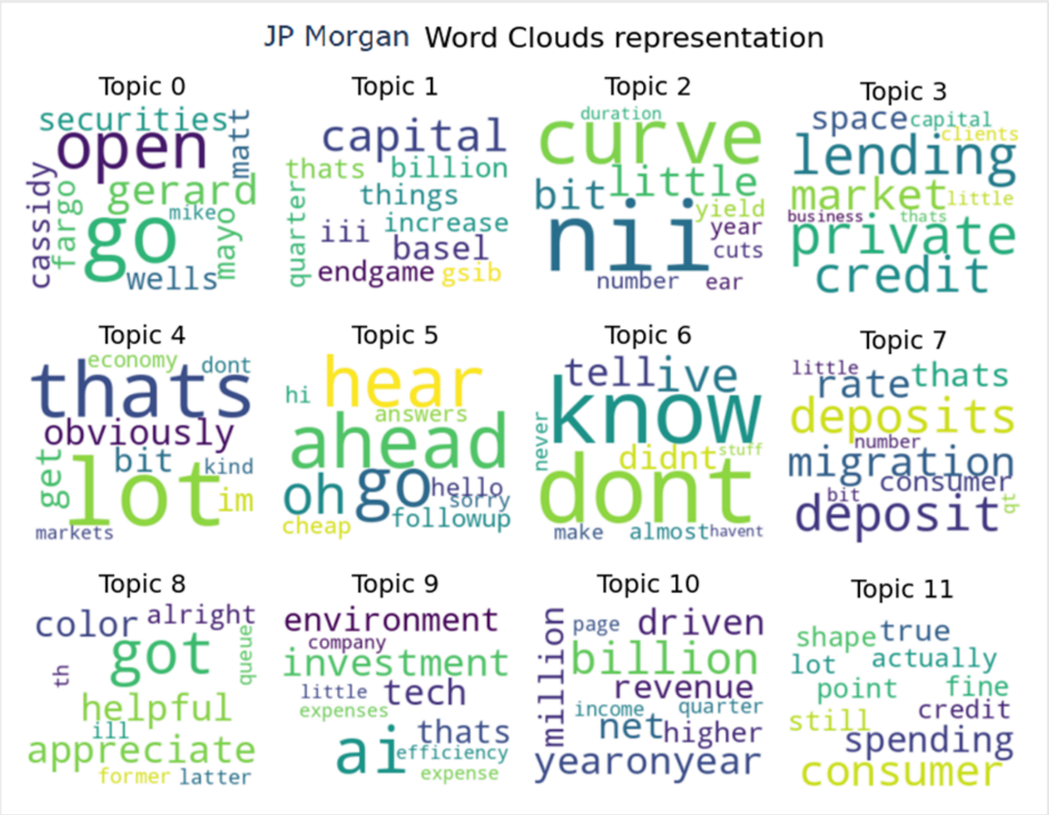


**Figure 2a:** JP Morgan Topic clouds.

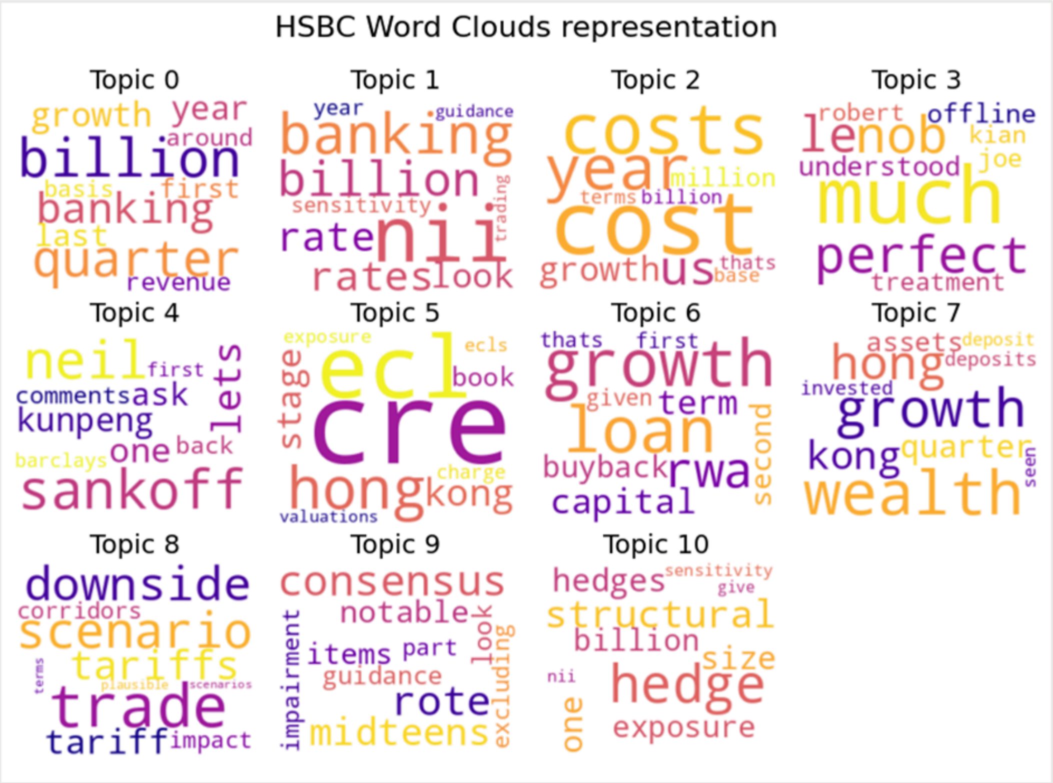
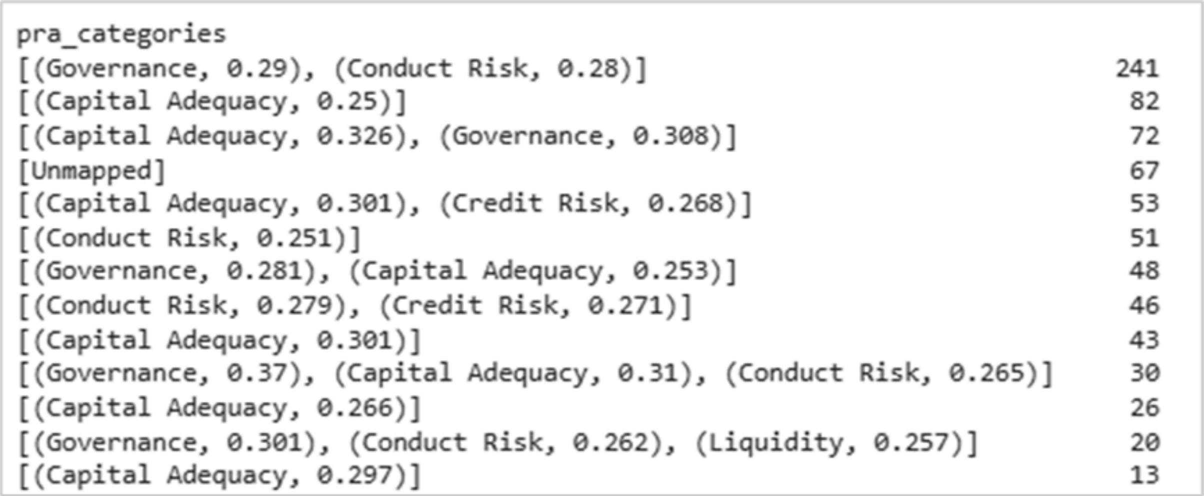
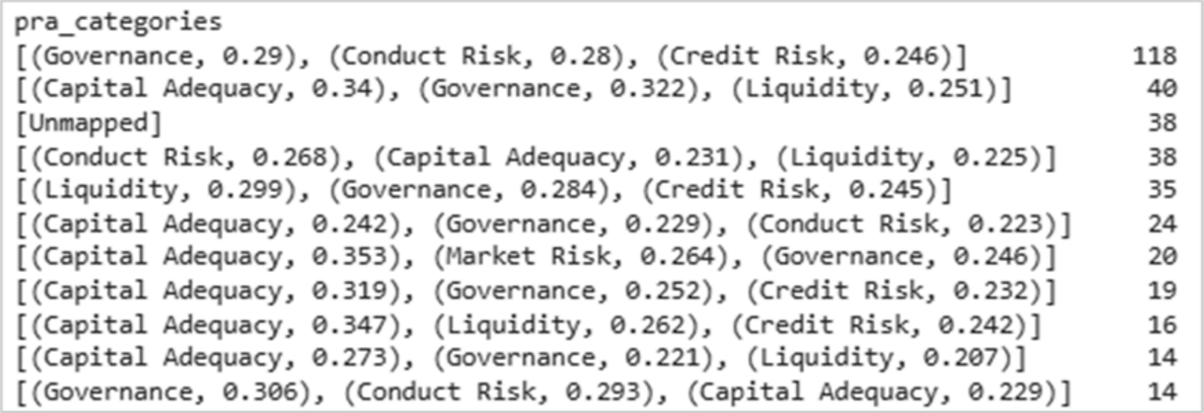


**Figure 2b:** HSBC Topic clouds.



```
pra_categories
[(Governance, 0.29), (Conduct Risk, 0.28)]                                      241
[(Capital Adequacy, 0.25)]                                                       82
[(Capital Adequacy, 0.326), (Governance, 0.308)]                                 72
[Unmapped]                                                                       67
[(Capital Adequacy, 0.301), (Credit Risk, 0.268)]                                53
[(Conduct Risk, 0.251)]                                                          51
[(Governance, 0.281), (Capital Adequacy, 0.253)]                                 48
[(Conduct Risk, 0.279), (Credit Risk, 0.271)]                                    46
[(Capital Adequacy, 0.301)]                                                      43
[(Governance, 0.37), (Capital Adequacy, 0.31), (Conduct Risk, 0.265)]            30
[(Capital Adequacy, 0.266)]                                                      26
[(Governance, 0.301), (Conduct Risk, 0.262), (Liquidity, 0.257)]                 20
[(Capital Adequacy, 0.297)]                                                      13
```

**Figure 3a:** JP Morgan PRA related themes.



```
pra_categories
[(Governance, 0.29), (Conduct Risk, 0.28), (Credit Risk, 0.246)]                118
[(Capital Adequacy, 0.34), (Governance, 0.322), (Liquidity, 0.251)]              40
[Unmapped]                                                                       38
[(Conduct Risk, 0.268), (Capital Adequacy, 0.231), (Liquidity, 0.225)]           38
[(Liquidity, 0.299), (Governance, 0.284), (Credit Risk, 0.245)]                  35
[(Capital Adequacy, 0.242), (Governance, 0.229), (Conduct Risk, 0.223)]          24
[(Capital Adequacy, 0.353), (Market Risk, 0.264), (Governance, 0.246)]           20
[(Capital Adequacy, 0.319), (Governance, 0.252), (Credit Risk, 0.232)]           19
[(Capital Adequacy, 0.347), (Liquidity, 0.262), (Credit Risk, 0.242)]            16
[(Capital Adequacy, 0.273), (Governance, 0.221), (Liquidity, 0.207)]             14
[(Governance, 0.306), (Conduct Risk, 0.293), (Capital Adequacy, 0.229)]          14
```

**Figure 3b:** HSBC PRA related themes.

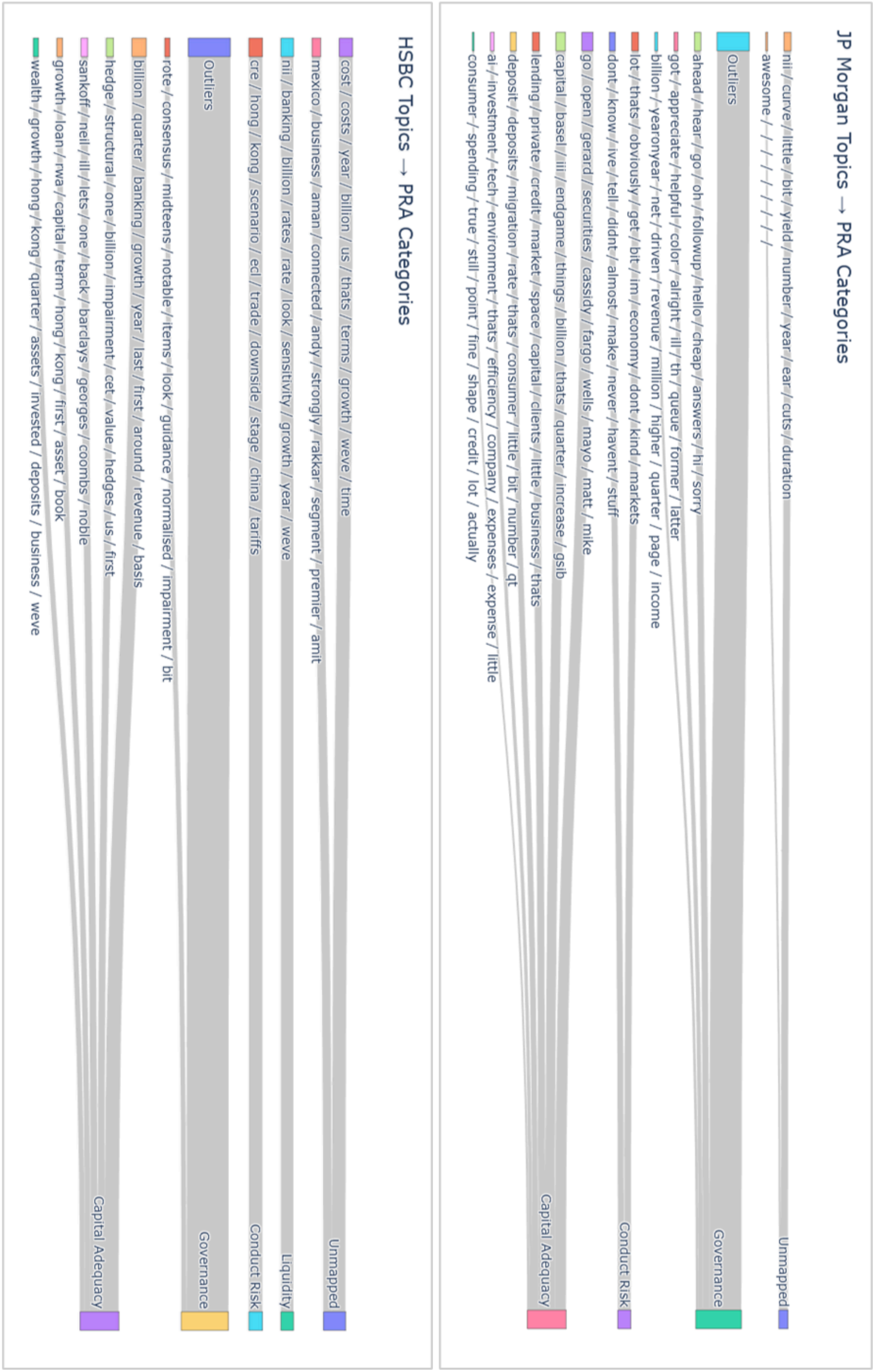

JP Morgan Topics → PRA Categories
nii / curve / little / bit / yield / number / year / ear / cuts / duration
awesome / / / / / / / / /
Outliers
ahead / hear / go / oh / followup / hello / cheap / answers / hi / sorry
got / appreciate / helpful / color / alright / ill / th / queue / former / latter
billion / yearonyear / net / driven / revenue / million / higher / quarter / page / income
lot / thats / obviously / get / bit / im / economy / dont / kind / markets
dont / know / ive / tell / didnt / almost / make / never / havent / stuff
go / open / gerard / securities / cassidy / fargo / wells / mayo / matt / mike
capital / basel / iii / endgame / things / billion / thats / quarter / increase / gsib
lending / private / credit / market / space / capital / clients / little / business / thats
deposit / deposits / migration / rate / thats / consumer / little / bit / number / qt
ai / investment / tech / environment / thats / efficiency / company / expenses / expense / little
consumer / spending / true / still / point / fine / shape / credit / lot / actually
Unmapped
Governance
Conduct Risk
Capital Adequacy
HSBC Topics → PRA Categories
cost / costs / year / billion / us / thats / terms / growth / weve / time
mexico / business / aman / connected / andy / strongly / rakkar / segment / premier / amit
nii / banking / billion / rates / rate / look / sensitivity / growth / year / weve
cre / hong / kong / scenario / ecl / trade / downside / stage / china / tariffs
Outliers
rote / consensus / midteens / notable / items / look / guidance / normalised / impairment / bit
billion / quarter / banking / growth / year / last / first / around / revenue / basis
hedge / structural / one / billion / impairment / cet / value / hedges / us / first
sankoff / neil / ill / lets / one / back / barclays / georges / coombs / noble
growth / loan / rwa / capital / term / hong / kong / first / asset / book
wealth / growth / hong / kong / quarter / assets / invested / deposits / business / weve
Unmapped
Liquidity
Conduct Risk
Governance
Capital Adequacy

**Figure 4:** HSBC & JP Morgan Topic to PRA flow diagram.

## 2.4. Sentiment Analysis

Four transformer models were initially assessed: two financial-domain FinBERT variants, CardiffNLP Twitter-RoBERTa, and DistilRoBERTa. The latter, fine-tuned on financial news, was selected as the most balanced in accuracy and generalisation across banking transcripts. The task was formulated as a 3-class sentiment classification (positive, neutral, negative) applied at the Q&A segment level.

FinBERT models provided strong financial alignment but struggled on smaller samples, while CardiffNLP handled informal tone yet lacked domain precision. DistilRoBERTa demonstrated the best trade-off between contextual understanding, domain fit, and computational efficiency, reaching 90% accuracy on manual validation and 0.91 F1 after fine-tuning.

Iterative improvements included class balancing, learning rate tuning, and early stopping. A stratified 80/20 train-test split was applied with weighted F1, accuracy, precision, and recall as evaluation metrics. Validation incorporated manual labelling for cross-checking sentiment reliability and quarterly consistency tracking.

The fine-tuned DistilRoBERTa achieved 92% accuracy and strong stability across 2023–2025 transcripts. Analysts expressed a more neutral tone, while bankers maintained consistently positive sentiment, most notably around capital adequacy and market risk (**Table 1**). These divergences indicate optimism bias in managerial communications versus analytical caution.

**Table 1:** Sentiment scoring by role.

| Role | Negative (%) | Neutral (%) | Positive (%) | Sentiment |
|---|---|---|---|---|
| Analysts | 4.5 | 86.8 | 8.6 | cautious |
| CFOs | 6.8 | 73.6 | 19.6 | optimistic |
| Executives | 6.8 | 84.3 | 8.9 | balanced |

In addition, JP Morgan exhibited greater sentiment variation (24.9%) than HSBC (13.4%) reflecting firm-specific differences in communication tone (**Table 2**).

**Table 2:** Cross-bank variation in sentiment and dominant topics.

| Bank | Model | Cross-bank Variation (%) | Top Topics |
|---|---|---|---|
| JPM | FinBERT (ProsusAI) | 24.9 | Market (8.4%)<br>Revenue (3.6%)<br>Regulatory (1.9%) |
| HSBC | FinBERT (Yiyanghkust) | 13.4 | Revenue (9.4%)<br>Market (8.1%)<br>Capital (2.3%) |

## 2.5. Evasion Detection

The evasion detection module aimed to maximise recall of evasive answers and generate a shortlist of flagged disclosures. Development began with a transparent rule-based baseline and progressed to NLI models (RoBERTa-large-MNLI, DeBERTa-MNLI, and zero-shot DeBERTa), where each Q&A pair was tested against direct and evasive hypotheses.

While heuristics were interpretable, they were rigid and NLI methods were limited by strict context windows. To improve context sensitivity, we experimented with RAG few-shot prompting, retrieving balanced exemplars using SBERT embeddings, though this provided limited performance gains.

Given the supervisory use case, threshold tuning of raw evasion scores was optimised to prioritise recall over precision, as missing evasive responses poses a greater supervisory risk than over-flagging. Validation used a stratified validation/test split of human-labelled transcripts, though manual labelling was time-consuming and models struggled to generalise consistently. Ensemble averaging of LLM outputs improved stability but varied by context.

DeBERTA was selected for binary labelling of answers as evasive or direct as it provided the best balance across metrics (evasive recall: 69%, F1 evasive: 0.28). In contrast, RoBERTa was selected for threshold free ranking (P@25%: 21.4%), with applications to produce shortlists of potentially evasive answers for review (**Table 3**).

**Table 3:** Model performances on JP. Morgan Test set.

| Model | Evasive Detected (%) | Evasive Missed (%) | Direct Detected (%) | Direct Missed (%) | P@25% (%) |
|---|---|---|---|---|---|
| Baseline | 93.8 | 6.2 | 12.6 | 87.4 | 10.7 |
| DeBERTa | 68.8 | 31.2 | 46.3 | 53.7 | 14.3 |
| Avg LLM | 62.5 | 37.5 | 38.9 | 61.1 | 17.9 |
| Blended (base + DeBERTa) | 75 | 25 | 19.7 | 82.1 | 17.9 |
| RoBERTa | - | - | - | - | 21.4 |

Evasiveness predictions for 2025 transcripts revealed that evasive behaviour clustered in capital adequacy for JP Morgan, and capital adequacy and costs & efficiency for HSBC (**Figure 5, 6**).

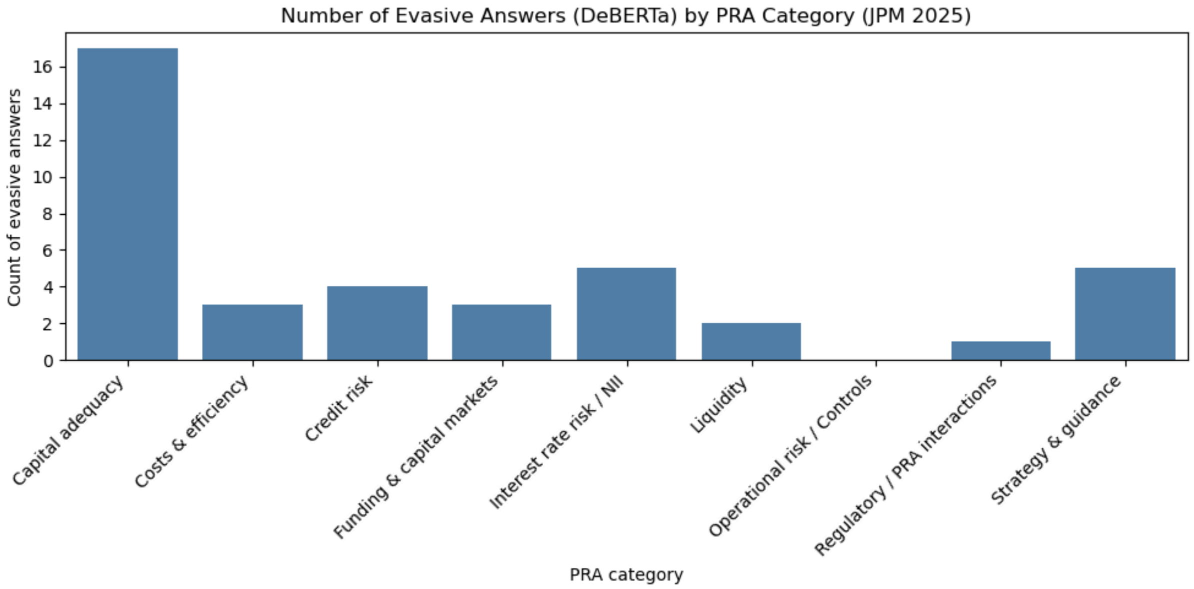


**Figure 5:** Evasive count by PRA category for JP Morgan Q&A 2025 transcripts.

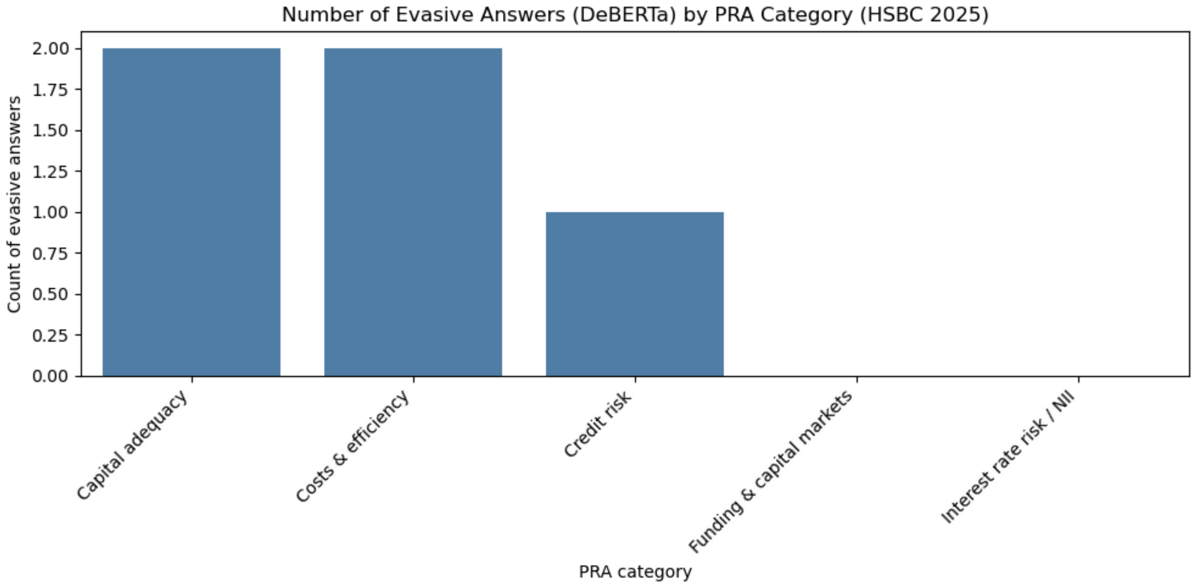


**Figure 6:** Evasive count by PRA category for HSBC Q&A 2025 transcripts.

By quarter, JPM evasiveness rose from Q1 to Q2, while HSBC showed fewer and declining evasive responses (**Figure** 7, **8**). These findings suggest supervisors should prioritise capital adequacy themes, and focus quarterly monitoring on JPM disclosures.

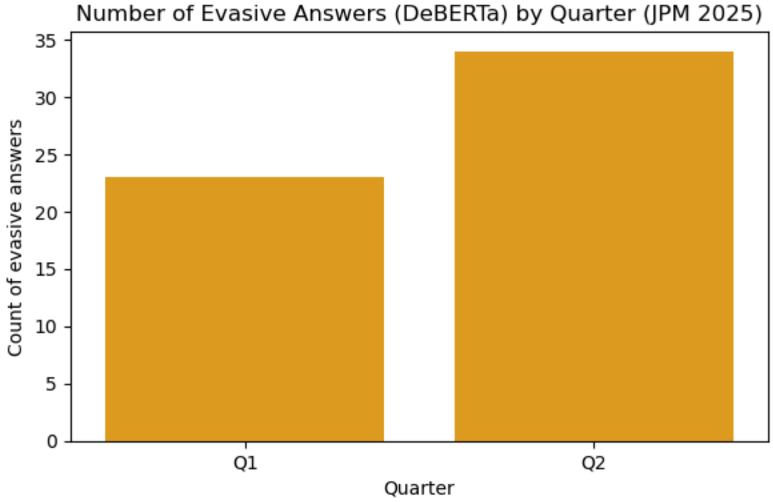


**Figure 7**: Number of evasive answers by quarter for JP Morgan Q&A 2025 transcripts.

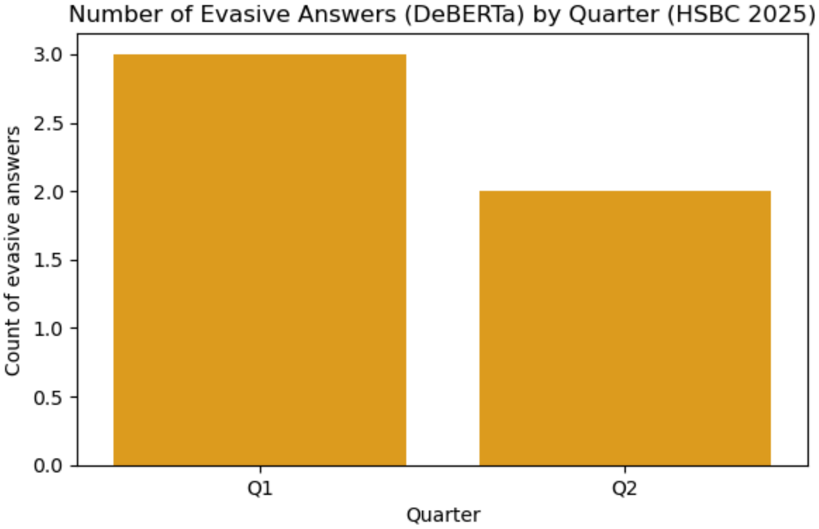


**Figure 8**: Number of evasive answers by quarter for HSBC Q&A 2025 transcripts.

### 2.6. Summarisation

The final pipeline stage consolidated outputs from topic modelling, sentiment analysis and evasion detection into concise, regulator-ready summaries. Using a large language model (GPT-4.1-nano), structured CSV outputs were synthesised into short, interpretable narratives aligned with PRA risk categories. The model outputs were designed to be concise and PRA-aligned, avoiding technical jargon to ensure interpretability for supervisory teams.

The summaries provided quarterly insights highlighting prevailing and emerging risks. In Q1-Q2 2025, analysts focused on capital adequacy and credit risk, while bankers emphasised strategy and operational resilience. Sentiment trends showed widening divergences between analyst caution and executive optimism, particularly on efficiency and cost management. Evasion scores added supervisory value by flagging less transparent responses, especially around liquidity and expenses.

Importantly, we extended the summaries into forward-looking "Q3 Watchouts," which distilled potential supervisory priorities such as pressure on net interest margins, operational evasiveness and credit risk exposures. This stage demonstrates how AI can transform complex transcript analysis into scalable PRA-aligned insights, supporting faster, data-driven oversight.

## 3. Conclusion & Recommendations

This project demonstrates how combining NLP techniques can transform unstructured Q&A transcripts into actionable supervisory insights. The pipeline revealed recurring areas of concern including capital adequacy, cost efficiency, and liquidity, where overly positive tone or evasive responses suggest where supervisory attention could be focused in future monitoring.

To strengthen confidence and scalability, the approach should be extended across more banks and time periods to test consistency and detect systemic trends. Enhanced labelling for evasive language and further fine-tuning of sentiment models on financial text would improve accuracy.

Ultimately, integrating this pipeline into a centralised PRA or BoE dashboard would enable continuous, data-driven monitoring of tone, sentiment and evasiveness across firms, supporting faster, more consistent and proactive supervisory review.

## 4. References

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J. and Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences*, 13(2), *p*.797. doi:https://doi.org/10.3390/app13020797