Employer Project: Final Report

Business Context and Objectives

The Prudential Regulation Authority (PRA) of the Bank of England plays a pivotal role in safeguarding financial stability in the UK by supervising banks, insurance, and investment firms. Global Systemically Important Banks (G-SIBs) are required to regularly report their capital and risk exposure, and typically provide quarterly press releases or announcements paired with published transcripts of earnings calls with financial analysts. While the announcements are thoroughly analysed by the PRA, information contained within the earnings call transcripts is currently underutilised. These transcripts could offer rich qualitative insights into banks' performance, sentiment shifts, and emerging risks. By integrating transcript analysis into regulatory frameworks, we aim to assess whether this methodology can enhance the PRA's risk identification and financial oversight capabilities.

This project evaluates the potential utility of earnings call transcripts for the PRA, with a focus on leveraging advanced Natural Language Processing (NLP) techniques and language models. Using JPMorganChase as a proof-of-concept, our analysis encompassed 14 quarters of transcripts to assess the feasibility of identifying emerging risks, financial volatilities, and discussions of regulatory impacts.

The key business questions this project aims to address are:

1. Can transcript analysis provide useful insights?
2. What emerging risks can be identified?
3. Can this approach be scaled for financial assessment of any banks' transcripts?

Project Development Process

1. Data Preparation and Initial Cleaning

The project commenced with the collection of JPMorganChase's earnings call transcripts spanning 14 quarters (2Q21–3Q24). Key attributes such as speaker names, titles, and the nature of the text (e.g., question, answer, or executive remarks) were extracted and organised in a tabular format, using a tailored script that leveraged various formatting features of the specific bank's transcripts. This was followed by manual merging of interrupted responses and removal of non-informative elements like greetings.

To contextualise the earnings calls, we also collected and tabularised four critical financial metrics reported by JPMorganChase over the 14 quarters:

- CET1 Capital Ratio, reflecting the bank's core capital adequacy
- Net Income, indicating profitability and financial health
- Earnings Per Share (EPS), showcasing per-share earnings
- Provision for Credit Losses, highlighting reserves for potential loan defaults.

Through this analysis, we identified quarter 1Q22 as a period reflecting particular financial strain that could be used to test the insights generated by various models (**Figure-1**). Subsequently, to identify emerging risks, we focused on the two most recent quarters (2Q24/3Q24), which were characterised by good financial performance.

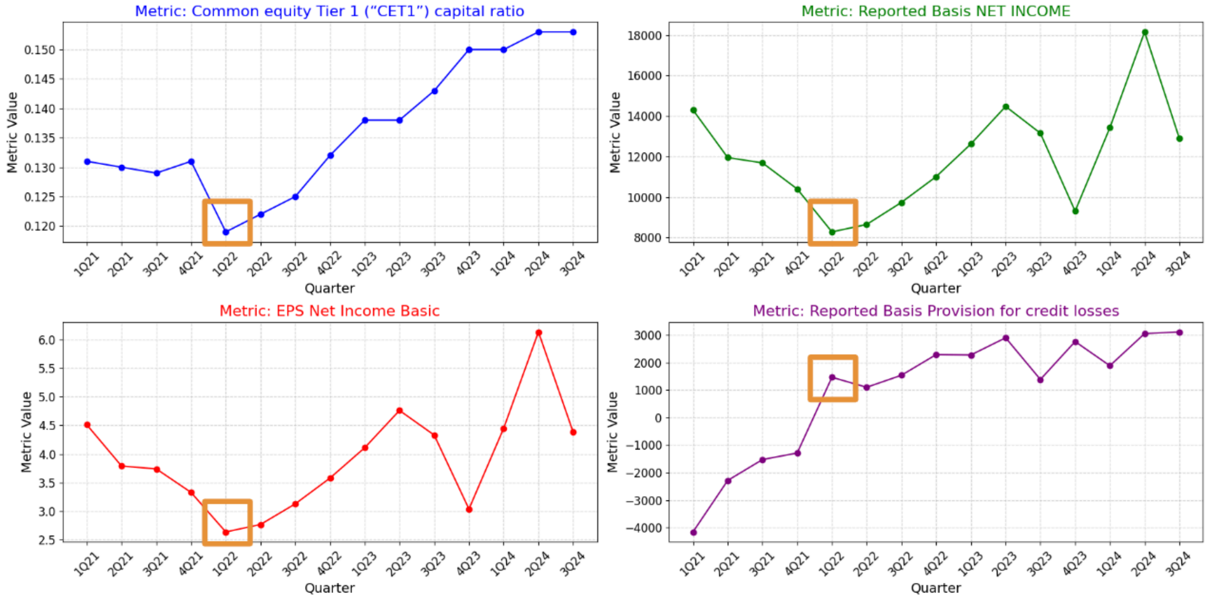

Figure-1: *Financial metrics from 1Q22 to 3Q24: CET1, Net Income, EPS, Provision for Credit Losses.*

2. Methodology

We explored and compared two broad sets of Natural Language Processing (NLP) approaches (**Figure-2**).

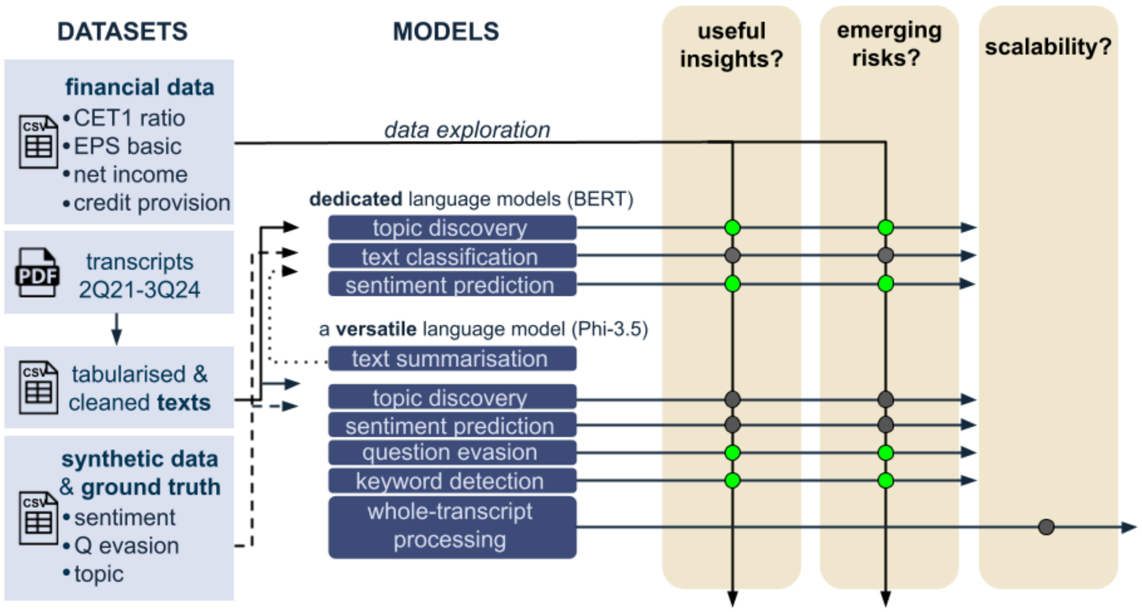

Figure-2: *Project development process. Green circles indicate models that form the recommended analysis pipeline.*

The first approach applied a series of dedicated, task-specific language models to the tasks they had been fine-tuned for. These included:

- **BERTopic**: an unsupervised topic discovery model designed to identify clusters of similar terms and group them into interpretable topics. Unlike models constrained by predefined categories, BERTopic provides flexibility to uncover emerging themes, making it a good candidate for exploring the variety of topics in earnings call transcripts.

- **FinBERT**: a BERT-based model fine-tuned on financial text. We explored two variants of FinBERT - one for text classification into 20 predefined financial topics (nickmuchi/finbert-tone-finetuned-finance-topic-classification from HuggingFace) and another for sentiment prediction (yiyanghkust/finbert-tone).

- **RoBERTa**: a general-purpose transformer model. Its variant (soleimanian/financial-roberta-large-sentiment) was employed for sentiment analysis. It was fine-tuned on a range of financial texts including earnings call transcripts, so we could expect it to perform well for our data.

The second approach was to use **Phi-3.5**, a versatile open-source language model on all the tasks mentioned above, as well as detection of keywords related to regulatory topics, such as Basel III, and question evasion. We also used it to generate concise summaries of transcript segments, which were tested as inputs for the task-specific models to assess the scalability of our pipeline.

3. Model Evaluation

To evaluate the performance of the NLP models on our data, two datasets were created to ensure robust testing:

- **Synthetic Data**: Text data with predefined sentiment, topic, and evasion labels was generated by GPT-4 and extensively refined manually to ensure accuracy. This dataset allowed us to test the models against controlled baselines and evaluate their precision in tasks with known outcomes.
- **Ground Truth Data**: Two transcripts' worth of texts were randomly sampled and manually annotated for sentiment, topic, and question evasion status. This dataset provided a benchmark for assessing the models' real-world performance.

Sentiment Analysis

To find the best model for sentiment classification, we tested FinBERT and RoBERTa on the manually labelled raw text inputs. RoBERTa consistently outperformed FinBERT, achieving an accuracy rate of 67%.

We then explored whether summarisation could enhance performance. However, we observed that it led to a higher proportion of texts being misclassified as neutral sentiment, ultimately reducing accuracy to 57%. This occurred because summarisation tends to preserve factual content rather than sentiment, causing the texts to sound more neutral overall.

Topic Discovery

To assess whether transcripts provide meaningful context to financial metrics, we tested two approaches of topic analysis: supervised topic classification with FinBERT and unsupervised topic discovery with BERTopic and Phi-3.5.

First, a variant of FinBERT fine-tuned on tweets was used to classify texts into 20 predefined topics, such as “company/product news”, “financials”, and “legal/regulation”. Tested on raw and Phi3.5-summarised text using its dedicated tokeniser, the model achieved ground truth accuracies of 12% and 26% respectively, outperforming the chance level of 5% (**Figure-3**). Notably, many errors involved closely related topics, with “legal\regulation” being misclassified as “fed/central banks” and “financials” mistaken for “macro” (**Figure-3**). Consequently, further exploratory analyses were conducted using topic probability distributions instead of model-assigned labels. However, the 20 predefined topics ultimately proved limiting, yielding insights that were sensible but crude at best.

Next, we turned to unsupervised topic modelling using BERTopic. BERTopic excelled at clustering broad themes in raw text, offering valuable insights into trends and recurring discussions within the transcripts. When BERTopic was applied to summarised text, the topics it discovered were more focused, as extraneous details were removed, allowing for clearer identification of underlying themes (**Figure-4**).

Question Evasion Detection

Detecting evasive responses in Q&A sessions required models to interpret the intent behind statements. Exploring the suitability of FinBERT for this task, topic probability distributions of questions and the corresponding answers were correlated and a range of correlation coefficient thresholds (0.65-0.9) were tested to categorise answers as evasive or non-evasive. Answers whose topic probability distributions were strongly correlated with those of the question (correlation coefficient >0.9) were considered non-evasive. This approach did not prove effective, achieving an accuracy of just 42% on synthetic data and a maximum of 50% accuracy on ground truth data (**Figure-3**).

Task performance for question-evasion improved to 62-67% accuracy by using Phi-3.5, well above the level of random chance (50%). However, in real-world scenarios, Phi-3.5 often overestimated evasiveness, frequently misclassifying short, incomplete or interrupted responses as evasive (**Figure-3**). While the task’s inherent subjectivity remained a challenge, Phi-3.5 provided valuable insights into evasive trends, enabling a deeper understanding of communication dynamics during analyst Q&A sessions.

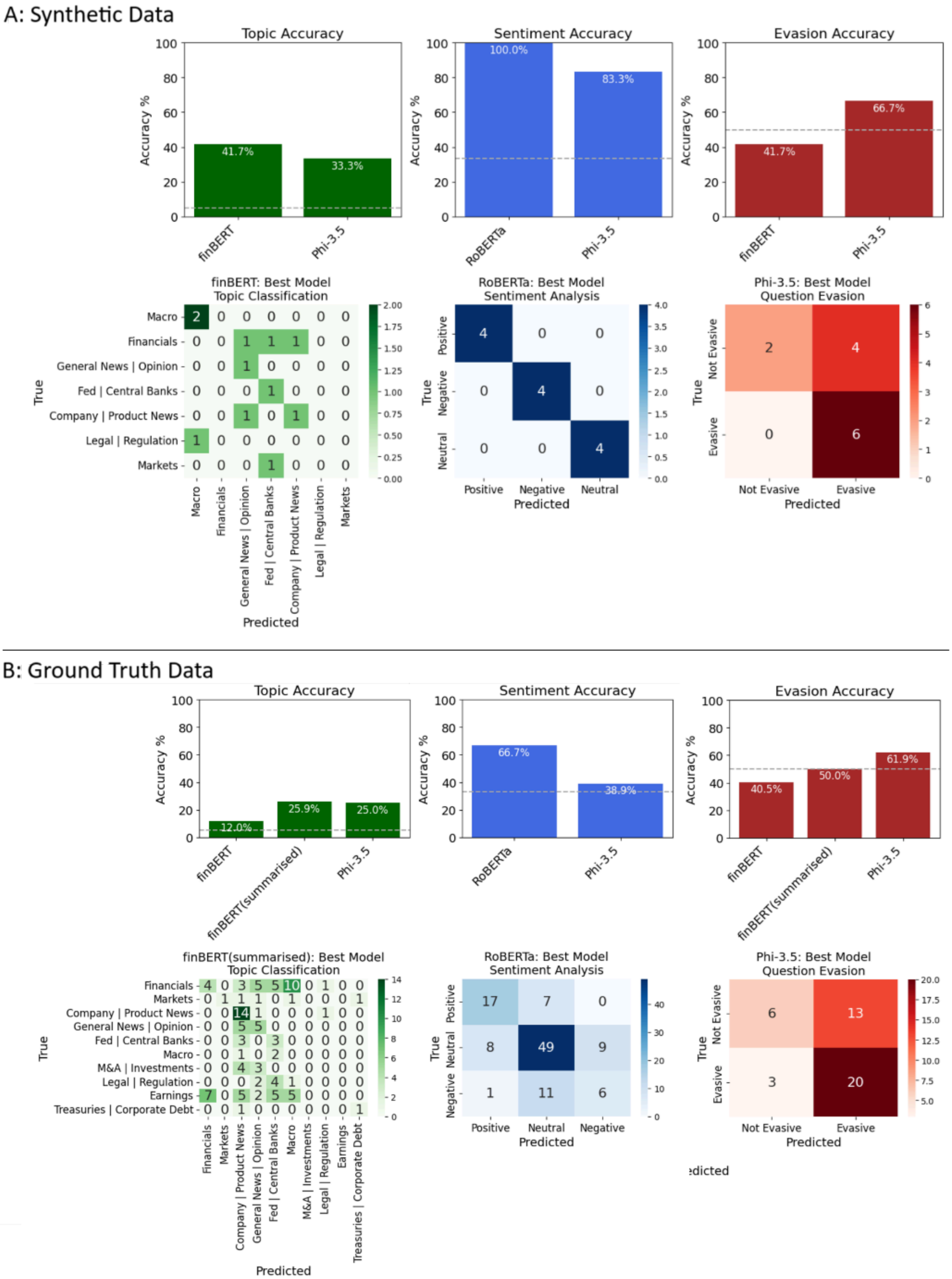

Figure-3: *Evaluation Dataset Accuracy results. Barcharts show accuracy of Topic, Sentiment and Evasion models against manually-labeled datasets, with dashed lines indicating level of random chance for label assignment. Confusion matrices are shown for the best performing model in each category.*

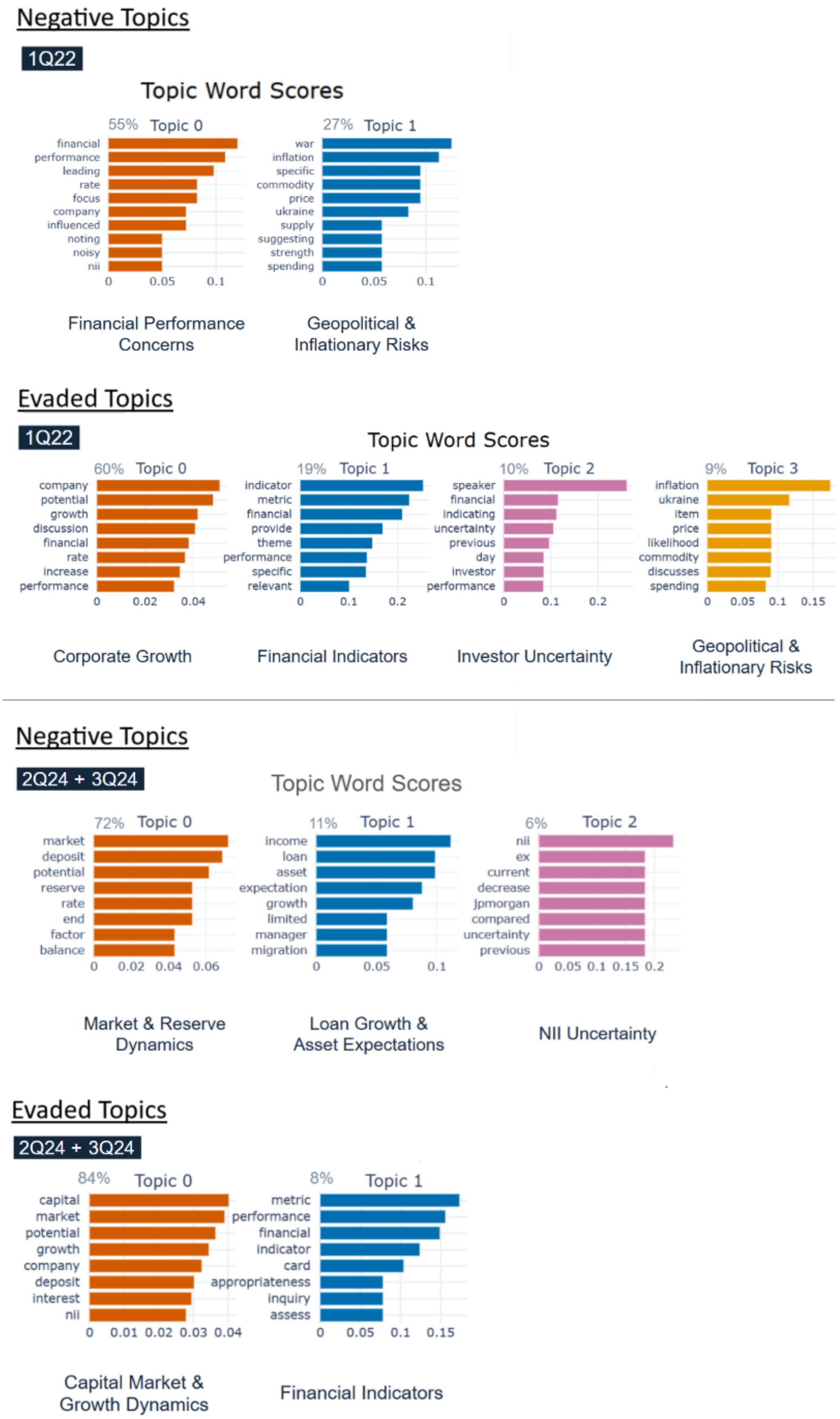

***Figure-4:** BERTopic-discovered topics in negative and evaded Q&As from 1Q22 (top panels) and from the two most recent quarters (bottom panels).*

Results

Can transcript analysis provide actionable insights?

To derive meaningful insight from transcripts, we recommend a solution pipeline that involves identification of texts with specific sentiment or question evasion status, followed by topic modelling using BERTopic on Phi-3.5-summarised text (**Figure-5**). By focusing on negative-sentiment Q&A in 1Q22, BERTopic highlighted themes like declining Net Interest Income in 55% of texts and geopolitical risks in 27% of texts (**Figure-4**). In evasive answers, the most common topic (60% of texts) concerned corporate growth, but geopolitical challenges were present too (9%). These findings enhanced our understanding of sentiment and thematic concerns during critical periods.

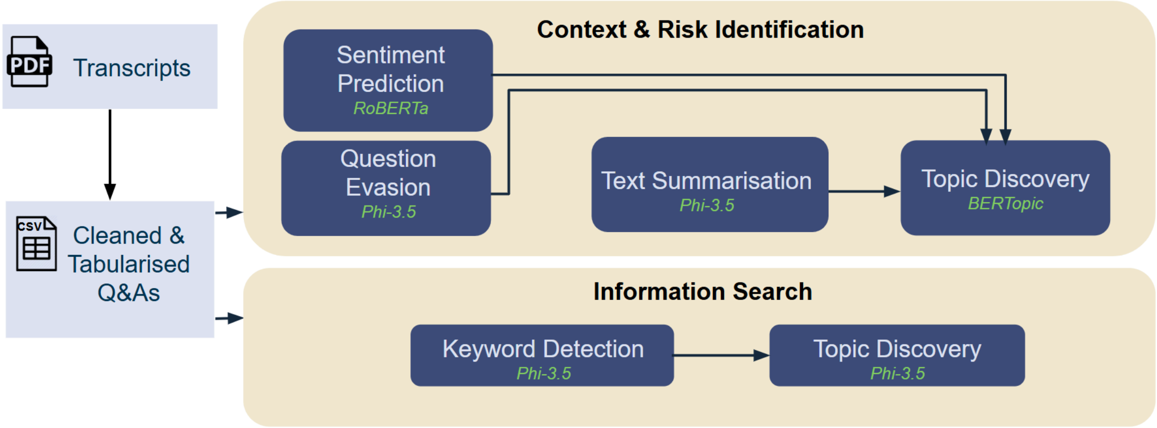

Figure-5: *Solution pipeline for earnings call transcript analysis.*

What emerging risks can be identified?

Emerging risks were identified through analysis of the two most recent quarters, where we observed a decline in the average sentiment, driven by greater neutrality at the expense of positive sentiment. Overlapping BERTopic-discovered themes in negative and evasive Q&As provided valuable insights into areas such as capital market uncertainty and reserve dynamics that might require closer regulatory attention (**Figure-4**).

Detection of keywords related to regulatory changes like Basel III revealed an association with topics such as capital adequacy and credit loss provisions. These discussions were categorised as evasive, but sentiment analysis revealed generally positive sentiments, offering context for actionable insights on emerging risks associated with the implementation of future regulations (**Figure-6**).

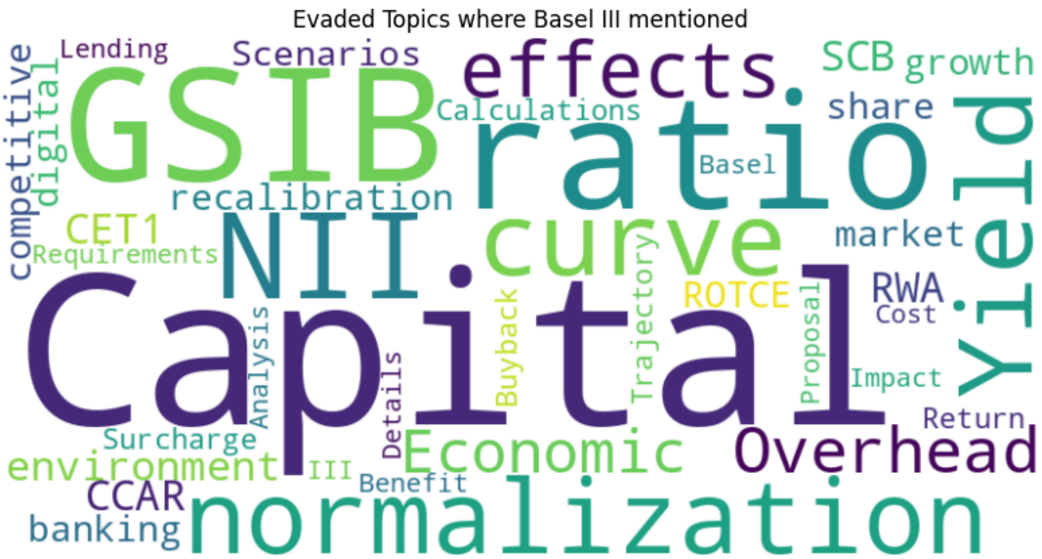

Figure-6: *Wordcloud of topics classified as evasive by Phi-3.5 in transcript discussions of Basel III. Size of word indicates relative frequency of occurrence.*

Can this approach scale to other banks?

The methodology shows potential for scalability with tailored preprocessing for different transcript formats. Automating transcript preprocessing would enhance adaptability. Generalised models, such as Phi-3.5 and BERTopic, already exhibit strong applicability across financial datasets, but fine-tuning of Phi-3.5 with domain-specific training datasets is likely to further improve performance.

Recommendations for Future Work

To improve scalability, it's recommended to automate preprocessing steps by using models like Phi-3.5 for decision making in merging interruptions and removing uninformative greetings. Fine-tuning of Phi-3.5 would also improve adaptability. Limitations, such as reduced sentiment accuracy with summarised text and overclassification of evasive responses, highlight the need for refinement, perhaps with longer summarisation texts. Exploring direct transcript analysis without preprocessing, using a versatile large language model, could further streamline the methodology and expand its application to systemic risk identification across the financial sector.