Bank of England

Employer project

RiskRadar Project Scope & Plan

GROUP 9

Overfit and Underpaid

Adeyinka Abdulrahman
Jessica A. Abreu
Alex Hamilton
Muhammad Latif
Rajen Lavingia
Arkadiusz Tomczak

Cambridge University
Career Accelerator, Data Science

2025

Table of Contents

Background & problem statement

Supervisors at the Bank of England face an overwhelming challenge: monitoring systemic risk across global systemically important banks (G-SIBs) requires analysing thousands of pages of unstructured disclosures quarterly, including earnings call transcripts, 10-Q/10-K filings, and analyst Q&A sessions.

Post-mortems on recent banking crises - including Silicon Valley Bank, Credit Suisse, and First Republic - reveal that warning signals often appear in public disclosures several quarters before collapse. These signals are subtle and buried within complex technical language, making them difficult to detect through manual review or traditional NLP methods.

The core problem is scale and subtlety. Supervisors must detect nuanced changes across dozens of institutions simultaneously. While quantitative metrics are easily monitored, qualitative indicators - such as CEO defensive language, answer avoidance patterns, or analysts' increasingly aggressive questioning - remain underutilised.

RiskRadar, our LLM-powered early warning system, transforms earnings materials into reliable supervisory signals. It employs multi sub-agent architecture combining topic modelling, sentiment analysis, answer avoidance detection, and novelty scoring to benchmark signals across G-SIBs.

Key questions we aim to address:

1. Do executives exhibit measurable linguistic stress patterns before adverse performance?
2. Can we detect answer avoidance on critical metrics (NIM/NII, deposit beta, CoR)?
3. What cost-performance trade-offs exist between models?
4. How can we ensure complete source traceability for regulatory compliance?

Team roles & ways of working

We adopt a hybrid approach combining primary ownership with collective responsibility, ensuring cross-functional expertise and resilience.

Primary Responsibilities:

Model Development Lead: *Alex Hamilton* and *Rajen Lavingia* - Leading LLM evaluation and multi-agent architecture implementation. Responsible for researching, creating prototypes, and testing LLMs and NLP models.

Evaluation and Metrics Lead: *Arkadiusz Tomczak* - Defining evaluation criteria (topic coherence ≥0.55, extraction F1 ≥0.85, citation faithfulness ≥95%). Creating comparison matrices and ensuring reproducibility.

Data Pipeline Lead: *Muhammad Latif* - Building of ingestion pipelines, processing transcripts and filings, and integrating models into the dashboard.

Project Oversight Lead: *Adeyinka Abdulrahman* - Coordinating timelines, tracking milestones, ensuring audit-ready outputs aligned to PRA requirements. Managing risk mitigation strategies.

Financial Insights Lead: *Jessica A. Abreu* - Bridging technical outputs with regulatory interpretation, defining KPI ontology (CET1, LCR/NSFR, NIM, CoR), and validating supervisory utility.

Collaborative Workflows:

We operate in 2-week sprints with mid-sprint check-ins and formal sprint reviews. Weekly synchronous meetings (30 minutes) accommodate time zone differences (UK±2h, scheduled 10:00-15:00 UK time). Asynchronous collaboration with shared documentation (Google Drive and GitHub).

Risk Mitigation:

Each component has secondary owners understanding implementation sufficiently to ensure continuity. Human-in-the-loop validation with PRA supervisors ensures

relevance. All outputs must be grounded with source citations following cite-or-abstain policy.

Project plan

The project follows an agile approach with three 2-week sprints, delivering incremental value while maintaining flexibility (Figure 1).

Sprint 1 (Weeks 1-2): Data Foundation & Model Baseline

- Establish data pipeline using earnings transcripts from JPMorgan, HSBC, Barclays, and Santander.
- Conduct systematic model evaluation comparing models and FinBERT across key metrics: accuracy in risk detection, processing speed, and API costs.
- Build the multi-agent framework, where specialised agents handle sentiment analysis, topic modelling, and metric extraction independently.
- Create document store with metadata and vector index using finance-specific embeddings.
- Conduct systematic model evaluation comparing GPT-4, Claude, Llama, FinBERT across metrics:
  - Accuracy in risk detection.
  - Processing speed and API costs.
  - Topic coherence (target ≥0.55).
  - Extraction F1 (target ≥0.85).

Build baseline multi-agent framework with specialised agents for sentiment analysis, topic modelling, and metric extraction

Deliverables: Baseline system, initial model comparison, functioning prototype

Sprint 2 (Weeks 3-4): Advanced Architecture & Answer Avoidance

- Implement modular multi-agent system with finite-state orchestration:
  - Sentiment tracking using FinBERT + Loughran-McDonald lexicons.
  - Topic modelling with BERTopic (prepared vs Q&A separately).

    - Metric extraction with KPI ontology.
    - Answer avoidance detection using semantic alignment and deflection scoring.
    - Novelty detection to flag firm-specific issues or sector trends.
- Develop Streamlit dashboard mockups for early feedback.
- Conduct Assignment 2 preliminary presentation.
- **Deliverables:** Integrated agents, answer avoidance scores, preliminary interface.

Sprint 3 (Weeks 5-6): Integration, Validation & Delivery

- Integrate components into complete RiskRadar system.
- Implement Supervisory Bank Condition Indicator (S-BCI): composite score (0-100) blending prudential metrics with text-derived signals.
- Validate against historical crises (COVID-19 Q1 2020, SVB 2023).
- Conduct QA with edge-case testing.
- Implement citation requirements (≥95% faithfulness).

Final deliverables:

- 1,500-word technical report with comprehensive methodology and findings.
- Interactive Streamlit dashboard featuring Answer-Avoidance Heatmap, S-BCI indicators, and peer benchmarking.
- Complete codebase on GitHub with version control and audit trails.
- Live demonstration to Bank of England stakeholders.

Project Management:

- Continuous issue tracking via GitHub.
- Weekly sprint reviews with documented decisions.
- PR reviews mandatory.

Approach

Our approach implements a grounded agentic pipeline with strict guardrails ensuring regulatory compliance.

Sprint 1 establishes data ingestion and benchmarks candidate models. Evaluation uses intrinsic metrics (topic coherence ≥0.55, extraction F1 ≥0.85, citation faithfulness

≥95%) and extrinsic utility (correlation with market stress indicators, time saved ≥30%). The baseline multi-agent framework implements specialised agents with narrow contracts and timeouts.

Sprint 2 optimises modules with domain-specific prompts. Answer avoidance detection employs slot coverage, question-answer semantic alignment, and hedging lexicons to compute avoidance scores. Novelty detection uses cosine similarity to identify topics not recently covered by peers. Behavioural stress detection incorporates hedge word frequency, response length patterns, and complexity shifts.

Sprint 3 validates end-to-end performance through back-testing against known stress periods and normal baselines. The Supervisory Bank Condition Indicator (S-BCI) provides a composite score decomposing capital resilience (25%), liquidity/funding (20%), credit cycle pressure (25%), market risk (10%), and text-derived signals (20%).

All outputs maintain complete source traceability through citation spans. Human-in-the-loop validation ensures interpretability. Risk mitigation includes bias testing, model interpretability checks, and positioning LLMs as summarisation aids rather than decision-making tools. The framework remains auditable, reproducible, and scalable for future expansion.

Appendix

Figure 1: *RiskRadar roadmap*