FSE 570 Capstone Project Proposal

Earnings Call & Risk Intelligence Engine
for Financial Decision Support

Team Members:

Name	Email
Shalin Nilesh Bhavsar	sbhavsa8@asu.edu
Ramanuj Magadi Anantha Krishna	rkrish79@asu.edu
Freya Hemalbhai Shah	fshah14@asu.edu
Kruthika Suresh	ksures21@asu.edu
Harihara Yashwanth Veldanda	hveldan1@asu.edu

Date: February 2026

Project Deadline: April 30, 2026

Arizona State University

Contents

1 Problem Statement (25%)

1.1 Real-World Problem Definition

Financial institutions and investment analysts face a critical challenge: **tracking how corporate risk narratives evolve over time across earnings calls and SEC filings**. Public companies disclose extensive risk information quarterly through earnings call transcripts and regulatory filings (10-K, 10-Q), yet this data presents severe analytical barriers:

1.1.1 Problem Requirements:

1. **Volume Overload:** A single S&P 500 company produces ∼100,000 words/quarter of risk-related text; tracking 20 companies = 8M words/year
2. **Manual Analysis Limitations:** Detecting subtle quarter-over-quarter changes in risk emphasis requires reading hundreds of documents sequentially
3. **Lack of Temporal Tracking:** No automated system tracks longitudinal changes in corporate risk communication (e.g., cybersecurity mentions increasing 340% over 3 years)
4. **Hallucination Risk:** General AI tools (ChatGPT, etc.) cannot enforce evidence grounding or cite specific sources, violating financial governance standards
5. **Integration Gap:** Earnings calls (conversational) and SEC filings (legal) are analyzed separately, missing critical discrepancies between management tone and formal disclosures

Current State: Analysts manually read documents, create ad-hoc spreadsheets, and miss early warning signals. Tools like Bloomberg Terminal provide document access but no automated longitudinal risk intelligence.

1.2 How This Project Solves the Problem

This project delivers an **AI-powered Earnings Call & Risk Intelligence Engine** that:

- **Automates longitudinal tracking:** Detects quarter-over-quarter and year-over-year changes in risk emphasis across 900+ documents
- **Enforces evidence grounding:** Every AI-generated insight includes citations to specific filing sections (no hallucinations)
- **Quantifies qualitative disclosures:** Transforms narrative text into measurable indicators (risk frequency, sentiment drift, uncertainty scores)
- **Integrates heterogeneous sources:** Compares management tone in earnings calls against formal risk disclosures in SEC filings
- **Provides explainable insights:** Generates audit-ready reports with source attribution

1.2.1 Value to Stakeholders:

- **Investment Analysts:** Reduce due diligence time from 40 hours → 2 hours per company; detect risk signals 2–4 quarters earlier
- **Credit Officers:** Identify borrower risk deterioration before financial metrics reflect problems (e.g., increasing caution in disclosures)
- **Corporate Strategy Teams:** Benchmark risk posture against competitors; track industry-wide responses to regulatory changes
- **Academic Researchers:** Enable large-scale analysis of corporate communication behavior (previously limited by manual analysis)

1.3 Societal & Industry Impact

Financial Markets Efficiency: Early detection of corporate risk changes improves capital allocation, reducing systemic risk from overlooked warnings (cf. 2008 financial crisis where risk disclosures signaled problems analysts missed).

Regulatory Transparency: Demonstrates how AI can enhance disclosure effectiveness, supporting SEC initiatives for improved investor protection.

Economic Impact: Better investment decisions reduce portfolio volatility; estimated 5–10% improvement in risk-adjusted returns = billions in value for institutional investors managing $30+ trillion in U.S. equities.

Research Contribution: Creates open-source methodology for temporal NLP analysis of financial disclosures, advancing computational finance and corporate governance research.

2 Data Sources (25%)

2.1 Dataset 1: SEC EDGAR Financial Filings (Primary Source)

Source	U.S. Securities and Exchange Commission EDGAR database
Access	`https://www.sec.gov/edgar` — API: `https://data.sec.gov/`
Legal Status	Public domain (U.S. government data), completely free, no restrictions
Size	500+ filings (20 companies × 25 filings/company over 5 years)

2.1.1 Structure:

- **Document Types:** 10-K (annual reports), 10-Q (quarterly reports)
- **Key Sections:** Item 1A (Risk Factors), Item 7 (MD&A – Management Discussion & Analysis)
- **Content Volume:** 30 million characters total; avg 35,000 chars/Risk Factors section
- **Time Coverage:** 2020–2024 (captures COVID, inflation, AI disruption, geopolitical risks)
- **Companies:** Apple, Microsoft, Google, Amazon, JPMorgan, Bank of America, Pfizer, Johnson & Johnson, Walmart, ExxonMobil (+ 10 more across tech, finance, healthcare, retail, energy sectors)

Relevance: Risk Factors sections are legally mandated comprehensive inventories of all material risks; longitudinal analysis reveals how corporate risk posture evolves with business conditions.

2.1.2 Heterogeneity:

- 5 industry sectors with distinct risk vocabularies (tech: IP/AI; finance: credit/market; healthcare: FDA/litigation)
- Annual (10-K) vs quarterly (10-Q) reporting cadences
- Formal legal language with embedded quantitative metrics
- Document length variation: 20,000–80,000 characters

2.2 Dataset 2: Public Earnings Call Transcripts (Secondary Source)

Source	Financial Modeling Prep API (primary) + Kaggle public datasets (backup)
Access	https://financialmodelingprep.com/api (free tier: 250 calls/day) — Kaggle: https://www.kaggle.com/datasets/tpotterer/motley-fool-scraped-earnings-call-transcripts
Legal Status	Publicly disclosed investor communications, licensed for research use
Size	400+ transcripts (20 companies $\times$ 20 quarters)

2.2.1 Structure:

- **Components:** Prepared management remarks (2,000–5,000 words) + Q&A session (3,000–6,000 words)
- **Content Volume:** 20 million characters total; avg 50,000 chars/transcript
- **Speakers:** CEO, CFO, division heads (management) + equity analysts (questioners)
- **Format:** Transcribed conversational speech with speaker labels

Relevance: Earnings calls reveal real-time management priorities and tone 30–45 days before formal SEC filings; Q&A sections expose analyst concerns that management may downplay in prepared remarks.

2.2.2 Heterogeneity:

- Conversational vs written text (verbal fillers, fragmented sentences, emotional cues)
- Multi-speaker dynamics (management optimism vs analyst skepticism)
- Quarterly cadence (higher temporal resolution than annual 10-Ks)
- Unstructured dialogue vs structured SEC sections

2.3 Data Integration & Challenges

Linkage: Common keys (ticker symbol, fiscal quarter/year) enable temporal alignment
Combined Dataset: 900+ documents, 50 million characters, 250MB storage

2.3.1 Challenges & Solutions:

1. **HTML Parsing Variability (SEC):** Non-standard formatting in older filings → Multi-pattern regex extraction + BeautifulSoup HTML cleaning
2. **Section Extraction Failures (∼2–5%):** Tables/custom layouts → Fallback to broader text capture + manual validation flags
3. **Transcription Errors (Calls):** Speech-to-text mistakes → Accept minor errors (focus on themes not exact wording) + cross-reference with SEC filings
4. **Missing Transcripts (∼15%):** Not all companies publish calls → Document coverage gaps, ensure 80% minimum threshold
5. **Temporal Misalignment:** Calls occur 30 days before filings → Explicitly model lag; analyze as leading indicator vs comprehensive record
6. **Boilerplate Text:** Standard legal disclaimers repeated across filings → Document common phrases; optional removal or model learns to ignore
7. **Rate Limiting (API):** 250 calls/day limit → Schedule collection over 2–3 days OR use Kaggle backup dataset

Data Accessibility: Both datasets are publicly available with no authentication barriers. SEC provides free bulk downloads; FMP offers free tier sufficient for project scope. No proprietary data or PII involved (executives mentioned in professional capacity only).

3 Methodology (30%)

3.1 Advanced Data Science Techniques

3.1.1 3.1 Natural Language Processing Pipeline

Technique 1: Financial Domain NLP (Transformers)

- **Model:** FinBERT (BERT fine-tuned on financial text) for sentiment analysis and risk classification
- **Justification:** Domain-specific language models outperform general models on financial terminology (0.85 vs 0.65 F1-score on financial sentiment benchmarks)
- **Application:**
  - Classify risks into categories (regulatory, cyber, market, operational, financial, reputational, technology)
  - Extract sentiment polarity (−1 to +1) and uncertainty scores (0–100)
  - Identify forward-looking statements vs historical discussion

Technique 2: Temporal Change Detection (Statistical NLP)

- **Methods:**
  - TF-IDF vectorization with cosine similarity for quarter-over-quarter comparison
  - Change-point detection algorithms (PELT, Bayesian online changepoint detection) to identify structural breaks in risk language
  - Named Entity Recognition (NER) for risk-related entities (regulations, competitors, technologies)
- **Justification:** Detects not just what risks exist but when they emerge, intensify, or diminish
- **Validation:** Ground truth from known events (e.g., COVID emergence in Q1 2020 should show pandemic risk spike)

3.1.2 3.2 Machine Learning Models

Model 1: Multi-Label Risk Classification

- **Algorithm:** Gradient Boosting (XGBoost) with TF-IDF features + BERT embeddings
- **Training Data:** Hand-labeled sample of 200 risk factor paragraphs (10 categories)
- **Features:** Word frequencies, n-grams, sentence embeddings, document metadata
- **Validation:** 80/20 train-test split, 5-fold cross-validation
- **Target Metric:** F1-score > 0.70 (benchmark: random baseline ~ 0.10)

Model 2: Sentiment Regression

- **Algorithm:** Fine-tuned FinBERT with regression head for continuous sentiment scores
- **Training:** Transfer learning from FinBERT weights + fine-tuning on Financial PhraseBank dataset
- **Validation:** RMSE on held-out test set; correlation with stock price movements (sanity check)

Model 3: Risk Emergence Detection

- **Algorithm:** Bayesian changepoint detection on time-series of keyword frequencies
- **Justification:** Probabilistic approach quantifies uncertainty in detecting new risks
- **Validation:** Precision/Recall on synthetically inserted "new risks" in historical data

3.1.3 3.3 Retrieval-Augmented Generation (RAG) System

Architecture:

- **Embedding Model:** sentence-transformers/all-mpnet-base-v2 (768-dim dense vectors)
- **Vector Database:** FAISS (Facebook AI Similarity Search) for ∼50M character corpus
- **LLM:** GPT-4 or Claude-3 (via API) with strict prompt engineering for citation enforcement
- **Retrieval:** Top-k ($k = 5$) semantically similar passages for each query

Evidence Grounding Mechanism:

```
User Query: "How has Apple's cybersecurity risk changed?"
    ↓
1. Embed query → retrieve relevant Risk Factor passages (2020-2024)
2. Construct LLM prompt: "Based ONLY on these excerpts: [passages],
   answer: [query]. MUST cite sources as [Company, Filing Date,
   Section]. If insufficient evidence, say so."
3. LLM generates answer with inline citations
4. Validate citations point to actual retrieved passages
5. Return answer + source documents for user verification
```

Validation:

- **Citation Accuracy:** 95% of statements must trace to actual source text (automated checking)
- **Hallucination Detection:** Red-team testing with unanswerable queries (should return "insufficient evidence")
- **User Study:** 10 finance professionals rate explanation quality (target: 4.5/5 satisfaction)

3.2 Workflow

```
Data Collection → Preprocessing → Feature Engineering → Model Training
→ RAG Integration → Validation → Deployment
     ↓                ↓                ↓                    ↓
  API calls       HTML clean       TF-IDF vectors        XGBoost fit
 (1-2 weeks)     Text norm       BERT embeddings        FinBERT tune
              Sentence split    NER extraction        Changepoint fit
                                                           ↓
                                                      FAISS index
                                                      LLM prompts
                                                      Citation check
                                                           ↓
                                                       Cross-val
                                                      Test set eval
                                                      User testing
                                                           ↓
                                                     Streamlit UI
                                                      (Week 10)
```

3.3 Model Validation Strategy

1. Quantitative Metrics:

- Risk Classification: F1-score, Precision, Recall (target: > 0.70)
- Sentiment: RMSE, MAE (target: < 0.15 on $[-1, 1]$ scale)
- Changepoint: Precision/Recall on known events (target: > 0.80)
- RAG: Citation accuracy (target: $> 95\%$), answer relevance (BLEU/ROUGE scores)

2. Qualitative Validation:

- Case studies: Deep-dive on 3 companies comparing model outputs to manual analyst reports
- Expert review: 2 finance professionals evaluate 20 random predictions
- Ablation studies: Test impact of removing each component (e.g., FinBERT vs generic BERT)

3. Temporal Validation:

- Walk-forward testing: Train on 2020–2022, test on 2023–2024
- Event-based validation: Measure detection lag for known risk events (e.g., Silicon Valley Bank collapse)

4 Proposal Presentation and Clarity (10%)

Organization: Structured in order: Problem → Data → Methods → Management (2 pages)
Conciseness: Essential elements only; detailed specs in separate 28-page data documentation
Clarity: Plain language for business impact; technical precision for methodology
Professional Formatting: Headers, tables, clear section breaks; GitHub repository will include complete documentation

5 Project Management (10%)

5.1 Timeline & Milestones (13 weeks: Feb 1 – Apr 30, 2026)

5.2 Risk Identification & Mitigation

Risk 1: Data Collection Delays (Probability: Medium, Impact: High)

- **Monitoring Metric:** Filings collected per day; API success rate
- **Threshold:** If $< 50\%$ of target collected by Week 1 end → trigger mitigation
- **Mitigation Strategy:**
  - Primary: Use Kaggle pre-scraped datasets (immediate access, no API limits)
  - Secondary: Extend collection into Week 3 by parallelizing with preprocessing

Week	Dates	Phase	Deliverables	Status Check Goals
1–2	Feb 1–12	Data Collection & Proposal	500 SEC filings + 400 transcripts collected, validated	**Proposal Due (Feb 12):** Project proposal submitted
3–5	Feb 13–Mar 5	Preprocessing & EDA	Cleaned dataset, preprocessing pipeline, EDA report	**Status Update 1 (Mar 5):** Processed data ready, identified 5+ temporal trends
6–8	Mar 6–26	NLP Model Development	Risk classifier (F1$>$ 0.70), sentiment analyzer, changepoint detector trained	All 3 models validated on test set, performance metrics documented
9–11	Mar 27–Apr 2	RAG System Development	FAISS index built, LLM integrated, citation system working	**Status Update 2 (Apr 2):** RAG system answers queries with $>95\%$ citation accuracy
12–13	Apr 3–23	Integration & Testing	Full system integrated, user interface (Streamlit), end-to-end testing	System runs end-to-end, demo ready
14	Apr 24–30	Documentation & Delivery	Technical docs, presentation, demo video, final report	**Final Submission (Apr 30):** Presentation & Report

- Contingency: Reduce scope to 10 companies instead of 20 (still meets "large dataset" requirement)

Risk 2: NLP Model Underperformance (Probability: Medium, Impact: Medium)

- **Monitoring Metric:** F1-score on validation set during training
- **Threshold:** If F1 < 0.60 after initial training $\rightarrow$ trigger mitigation
- **Mitigation Strategy:**
  - Primary: Switch to proven FinBERT (already trained on financial text) vs training from scratch
  - Secondary: Simplify to rule-based classification + sentiment lexicons (LoughranMcDonald dictionary)
  - Contingency: Focus on descriptive analytics (keyword trends) vs predictive models; still valuable for longitudinal tracking

Risk 3: RAG Citation Accuracy Issues (Probability: Low, Impact: High)

- **Monitoring Metric:** % of LLM responses with verifiable citations; hallucination rate in testing
- **Threshold:** If citation accuracy $< 85\%$ $\rightarrow$ trigger mitigation
- **Mitigation Strategy:**
  - Primary: Implement strict prompt engineering with citation format enforcement + post-processing validation

- Secondary: Add deterministic keyword-based retrieval as fallback (less sophisticated but 100% traceable)
- Contingency: Simplify to retrieval-only system (no generation), just surface relevant passages for analyst review

5.3 Risk Monitoring Dashboard (Weekly Tracking)

Metric	Target	Current	Status
Data Collection Progress	100% by W2	TBD	Green/Yellow/Red
Model F1-Score	> 0.70	TBD	Green/Yellow/Red
RAG Citation Accuracy	$> 95\%$	TBD	Green/Yellow/Red
Timeline Adherence	On schedule	TBD	Green/Yellow/Red

Proactive Risk Management:

- **Daily standups (self-check):** 15-min review of progress vs plan
- **Weekly status reports:** Document metrics, blockers, adjustments
- **Bi-weekly advisor check-ins:** Review risks, get guidance on mitigation decisions
- **Contingency buffer:** Weeks 12–14 have slack for addressing unforeseen issues

6 Expected Deliverables (April 30, 2026)

1. **Working AI System:** Deployed locally or cloud (Streamlit interface)
2. **Technical Documentation:** 30+ pages (architecture, data pipeline, model specs, evaluation)
3. **User Guide:** Instructions for analysts to use the system
4. **Code Repository:** GitHub with clean, documented code + README
5. **Final Report:** 15–20 pages (problem, methods, results, limitations, future work)
6. **Presentation:** 25–30 slides + 10-min demo video
7. **Dataset & Models:** Processed data + trained model weights (for reproducibility)

6.1 Success Criteria:

- All 3 NLP models meet performance targets (F1> 0.70, RMSE< 0.15, Precision> 0.80)
- RAG system achieves $> 95\%$ citation accuracy on test queries
- User study shows 4.5/5 satisfaction from finance professionals
- System processes 5-year company history in < 5 minutes
- Complete documentation enables reproduction by other researchers