Team Name: *Team Oneironauts*

Project Title: Contextualized Financial Intelligence: A RAG-Based Approach to Financial Document Analysis

1. Motivation & Problem Statement

Analyzing financial documents like SEC filings is critical for investors, analysts, auditors, and regulators. These filings (e.g., 10-K, 10-Q, 8-K) contain valuable disclosures on a company's financial health, risks, executive decisions, and future plans. However, they are often lengthy, technical, and difficult to navigate.

While Large Language Models (LLMs) show impressive generative capabilities, they frequently hallucinate or misinterpret facts when applied to specialized domains such as finance. These models often lack the up-to-date, domain-specific grounding necessary for high-stakes tasks.

To address this, we propose a **contextualized Retrieval-Augmented Generation (RAG) system** focused on **financial document analysis**. Our system first retrieves relevant content from SEC filings and then generates answers, summaries, comparisons, or policy suggestions grounded in those disclosures. This aims to reduce hallucinations, improve factual accuracy, and support multiple analytical use cases.

2. What Do You Want to Change or Improve?

We aim to improve the **reliability**, **faithfulness**, and **domain awareness** of LLM outputs in financial contexts by:

- **Building a dedicated document index** using recent SEC filings, focusing on structure-preserving text and table representations.
- **Implementing a document-aware retrieval layer** that selects the most relevant semantic chunks.
- **Enhancing chunking** strategies by incorporating context-aware segmentation (e.g., based on document headings, semantic units, or table boundaries)
- **Reducing hallucinations** by tightly coupling answer generation with latest retrieved evidence from filings.
- **Supporting a broader range of financial analysis tasks** beyond question answering, such as summarization and comparisons etc.
- **Designing the system to be modular and extensible,** enabling future integration of real-time data(e.g., news, social media) or multimodal sources.

3. How Will You Measure Success?

We will evaluate both retrieval and generation components using different metrics

- **Generation Quality**

    - **F1 Score**: Overlap between generated answer and ground truth
    - **ROUGE-L**: For matching longest common subsequences
- **Retrieval Quality**
    - **Metric**: **Precision@k** and **Recall@k** (e.g., whether key supporting passages were retrieved in top-k)

4. What Data Will You Use?

- **Retrieval Corpus**
    - **Source:** US SEC EDGAR database
    - **Filing Types: 10-K**, **10-Q**, 8-K
        - **10-K:** Annual Report
        - **10-Q:** Quarterly Report
        - 8-K: Current Report
    - **Companies:** A curated list of 6–12 major public companies (e.g., Apple, Microsoft, Amazon, Johnson & Johnson etc)
- **Evaluation Dataset:**
    - We will build a **custom test set** comprising:
        - ~50–100 **task prompts** across multiple types:
            - Fact-based queries (e.g., “What was Amazon's net income in Q2 2023?”)
            - Executive compensation comparisons
            - Risk factor summarization
            - Market segment or revenue breakdowns
        - Each test set will include:
            - **User input question**
            - **Ground truth reference answer,** manually extracted from the source filings