**Date of proposal: 14/9/2025**

**Project Title:** **FinSight:** Intelligent Stock Prediction and Advisory Platform

**Group ID (As Enrolled in Canvas Class Groups): Group 6**

**Group Members (name , Student ID):**
**Huo Yiming A0328696J**
**Li Jiajun A0326795M**
**Samarth Soni A0329960U**
**SU Yuxuan A0329926N**
**Wang Yixi A0328469M**

**Sponsor/Client:** *(Company Name, Address and Contact Name, Email, if any)*
ArthAlpha
#2053 Prestige White Meadows, Whitefield, Bangalore, India 560066
Hunarpreet Singh
hunar@arthalpha.in

**Background/Aims/Objectives:**

1. *Background*

With the rapid growth of financial markets and the increasing availability of digital information, investors are often overwhelmed by massive volumes of stock data, financial reports, and market news. Traditional investment platforms in India and globally often lack personalized recommendations and dynamic adaptation to investor preferences.

This project aims to address that gap by developing an intelligent stock research and recommendation system that integrates financial fundamentals, user profiles, news sentiment, and predictive analytics.

We are glad to share that we have secured a **collaboration opportunity with ArthAlpha**, a SEBI-registered quant investment firm in India. They are providing our team with access to industry data and basic domain guidance. In return, they expect us to dedicate our efforts during the **September 25 – October 25 window** to focus on their problem statement: building an **AI Financial Research Analyst** that can analyze company fundamentals and provide structured investment insights using LLMs.

ArthAlpha has set two conditions:

1. We cannot share specific code, datasets, or hyperparameter details publicly (e.g., GitHub), but we can present results and describe general methods.
   We must treat October as a full-time academic project collaboration, aligning it with both our **Intelligent Reasoning Systems** and **Pattern Recognition** courses.

This collaboration not only grounds our project in industry relevance but also offers us exposure to real-world financial AI applications.

2. *Aims*

The aim of this project is to design and implement a **web-based stock research and prediction platform** that delivers personalized investment recommendations. This will combine financial fundamentals with cutting-edge AI methods including **large language models (LLMs), retrieval-augmented generation (RAG), deep learning, natural language processing, and vector-based similarity matching**.

3. *Objectives*

a) Develop a multi-dimensional user profile vector to represent investor preferences in sectors, risk tolerance, and thematic interests.
b) Integrate real-time stock and financial data, news articles, and regulatory filings into one web interface.
c) Implement a news browsing and analysis module with topic tagging and sentiment extraction, feeding dynamically into user profiles.
d) Build a stock trend prediction module using time-series modeling and machine learning.
e) Design an AI Financial Research Analyst module (ArthAlpha collaboration focus) that can:
- Parse company financial reports and market data
- Generate structured research outputs (ratios, valuations, peer comparisons)
- Provide evidence-backed recommendations and risk assessments
- Incorporate both quantitative and qualitative reasoning

g) Deliver a responsive front-end and scalable back-end system to ensure smooth real-time interaction.

**Project Descriptions:**

# 1. Market Research

1.1 Industry Trends

The global fintech market was valued at USD 340.10 billion in 2024. The market is projected to be worth USD 394.88 billion in 2025 and reach USD 1,126.64 billion by 2032, exhibiting a CAGR of 16.2% during the forecast period.Investors increasingly demand AI-driven solutions to filter vast financial data and make informed decisions

1.2 Competitive Landscape

Existing platforms such as Robinhood, Betterment, and Eastmoney provide trading tools, ETF-based advisory, or information aggregation. However, they lack dynamic user profiling, news sentiment integration, and AI-powered explainable personalized recommendations.

1.3 User Requirement

Retail investors struggle with information overload and lack professional analysis tools. Intermediate investors seek customizable recommendations that reflect their risk tolerance and sector preferences. They really need intelligent trading tools to guide their investment.

# 2. Project Scope

Data Scope: User Profile\Stock Data\News
Module Scope: News Browsing Module\Stock Trend Prediction\Portfolio Recommendation
Architecture Scope:Web pages based on front-end and back-end

# 3. Data Collection and Preparation

At the outset, we acknowledge that privacy and security, compliance, and ethical sourcing will be important; also making sure we respect copyright, data licensing especially for non-public / proprietary data.

## 3.1 **User Profiles**

User Profiles is the core data structure for storing 32-dimensional user profiles, every user gets his own user profile which contains these dimensions:

1. Industry preferences (8-d)
2. Risk tolerance (1-d)
3. Investment time horizon preferences (1-d)
4. Theme preferences (1-d):such as AI, new energy, semiconductors, ESG, etc.

5. Location(1-d)
6. Factor investing (8-d): Market Size, Value, Quality, Momentum, Low Volatility, Growth, Dividend yield, ESG preference
7. Implicit dimensions (12-d): User click/stay/favorite/negative feedback behaviors construct a user behavior latent space (12 dimensions)

## 3.2 News

Our project collects three complementary News resources—market prices and fundamentals, news, and corporate disclosures. News processing follows a clear flow: **fetch → normalize → deduplicate → enrich → vectorize → store**. Besides detailed news information, each news can be abstracted as a vector which contains these dimensions (same as user profile):

1. Industry preferences (8-d)
2. Risk tolerance (1-d)
3. Investment time horizon preferences (1-d)
4. Theme preferences (1-d):such as AI, new energy, semiconductors, ESG, etc.
5. Location(1-d)
6. Factor investing (8-d): Market Size, Value, Quality, Momentum, Low Volatility, Growth, Dividend yield, ESG preference
7. Implicit dimensions (12-d): User click/stay/favorite/negative feedback behaviors construct a user behavior latent space (12 dimensions)

## 3.3 Stocks

The stock data acquisition and processing pipeline is designed to support real-time recommendation, trend prediction, and detailed visualization for end users.

### 3.3.1 Data Acquisition

Stock data will be collected from multiple reliable financial data APIs to ensure comprehensiveness and accuracy. Primary sources include Polygon.io for real-time and historical price data, fundamentals, and corporate events. Data will be fetched periodically via scheduled jobs using Apache Airflow, with incremental updates to minimize API calls and ensure data freshness.

### 3.3.2 Data Processing and Feature Engineering

Raw stock data will undergo cleaning to handle missing values, outliers, and timestamp alignment, followed by normalization to standardize numerical features. The feature extraction process incorporates technical indicators, fundamental metrics, and sentiment-aggregated features derived from news sentiment scores.

#### 3.3.3 Stock Vector Representation

Each stock will be represented as a 32-dimensional vector to enable similarity matching with user profiles. The vector is constructed by reducing dimensionality of structured financial features and encoded textual content from company descriptions and filings. Each vector contains these dimensions (same as user profile):

1. Industry preferences (8-d)
2. Risk tolerance (1-d)
3. Investment time horizon preferences (1-d)
4. Theme preferences (1-d):such as AI, new energy, semiconductors, ESG, etc.
5. Location(1-d)
8. Factor investing (8-d): Market Size, Value, Quality, Momentum, Low Volatility, Growth, Dividend yield, ESG preference
9. Implicit dimensions (12-d): User click/stay/favorite/negative feedback behaviors construct a user behavior latent space (12 dimensions)

# 4. System Design

## 4.1 News Browsing Module

Our news module is designed to display recent related news and update the user profile based on historical records.News articles from trusted financial sources are continuously ingested, cleaned, deduplicated, and tagged with metadata. Each news item is encoded into a 32-dimensional vector representation aligned with the user profile space.

When a user requests the news feed, the system retrieves relevant news vectors, applies similarity matching against the user's profile, and re-ranks results considering recency, diversity, and source quality. User interactions such as clicks, dwell time, saves, and dislikes are collected as feedback signals. These signals are then used to adjust the user profile vector incrementally, strengthening alignment with topics and factors of interest while down-weighting irrelevant content.

This creates a feedback loop: **news presentation → user behavior → profile update → refined recommendations**

## 4.2 Stock Trend Module

### 4.2.1 Recommendation (Content-Based Matching)

This subsystem is designed to generate a personalized list of stock recommendations for each user by performing a semantic similarity match between their user profile vector and the vector representation of each stock. The core methodology is content-based filtering,

leveraging high-dimensional vector space search.

**a) Core Matching Algorithm**

First metric: Cosine Similarity between the user's profile vector and each stock's vector. This metric is chosen for its ability to measure orientation similarity independent of magnitude, making it ideal for comparing investment preference profiles.

Upon request, the system will retrieve the user's latest profile vector. It will then execute a K-Nearest Neighbors (KNN) search against the universe of stock vectors to identify the candidate pool with the highest cosine similarity to the user.

**b) Result Diversification**

We will implement the Maximal Marginal Relevance (MMR) algorithm. MMR optimizes the list by iteratively selecting stocks that balance relevance with diversity.

**c) Score Enhancement**

Sentiment Integration: To incorporate short-term market dynamics alongside long-term thematic alignment, the final ranking score will be a weighted blend of the semantic similarity score and a recent sentiment signal.

### 4.2.2 Trend Prediction

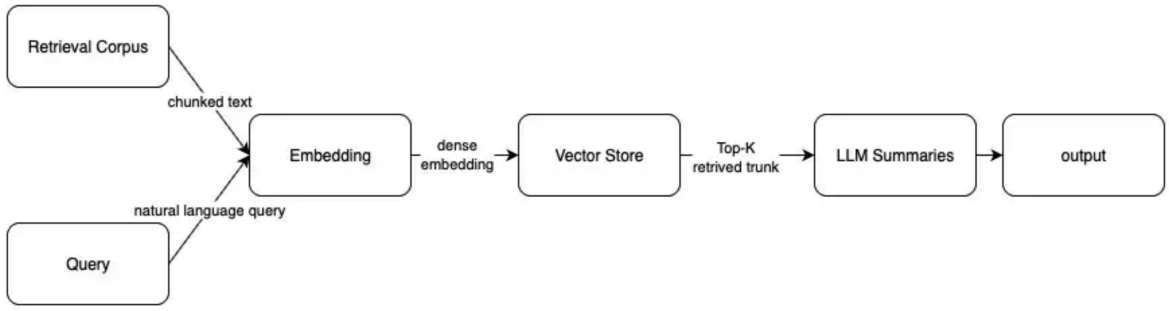
This subsystem combines a statistical forecaster (Prophet) with a neural sequence model (LSTM) to predict short-horizon stock trends. Prophet captures trend/seasonality and known regressors, while LSTM models nonlinear, regime-dependent residual patterns. Outputs are blended and then integrated with the recommendation scores before diversification (MMR)**.**

**a) Core Modeling Algorithm:** To predict 7-day / 14-day forward movement, this subsystem adopts a residual–hybrid forecaster that marries Prophet with an LSTM.

**b) Result Diversification:** The model outputs per-stock scores. Convert to a ranked list and apply the existing MMR to avoid sector/theme clustering.

**c) Score Enhancement**: Blend the hybrid forecast with your existing personalization and sentiment.

### 4.2.3 RAG for Details/Q&A




1. Retrieval Corpus: Company 10-K/10-Q, earnings call transcripts, news.
2. Embedding: Open AI text-embedding
3. Base LLM: Qwen2.5-32B generates summaries based
4. Vector Base: PostgreSQL with pgvector
5. Q&A Pipeline: Query → Embed → Top–k documents → Conclusion → Answer

## 4.3 Combined Investment Recommendation Module

**This module will form a core part of our application. It will function as** a semi-automated/AI-augmented **Financial Research analyst (FRA). A FRA is a sector-specific or stock-specific expert who goes deep into their area/stock of expertise and tries to analyse how the sector or company will perform in the short-term and mid-term future. Their reports help investors make decisions.**

**The main aim of our** "Combined Investment Recommendation Module", hereafter "CIRM" **is to give all its research on any stock or any sector based on publicly available data like company reports, past experience, and domain knowledge, news, etc. We will then compare it with other analysts' predictions to see how it has performed and to improve it. The module should also be able to revise its prediction based on news.** "Combined" refers to quantitative + qualitative, multiple data sources, etc.

**In brief,** Its purpose is to analyze stocks or sectors listed on Indian exchanges (NSE / BSE), aggregate data sources, run quantitative and qualitative analysis, and produce strategic reports and investment-recommendations (risk/return / scenarios) that are transparent, defensible, and updatable.

Our goal will be enabling an LLM in performing sentiment analysis and extracting signals from unstructured data like earnings calls, market news, and corporate filings.

This section explains how the CIRM will be structured (conceptually), the modules, workflows, and how things connect, in a high-level way without overpromising.

1. **Module architecture**
    - **Analytics & Metrics Module**: compute quantitative metrics, risk metrics, scenario modelling.
    - **Qualitative / Text Analysis Module**: process news, disclosures, risks,

sentiment; extract risk factors, regulatory events.
    - **Decision / Recommendation Module**: integrates quantitative & qualitative insights, produces structured output.
    - **Reporting & Visualization Module**: generate reports with data, charts, assumptions, confidence levels.
2. **Workflow**
    - Regular update cycle (say, quarterly + event-driven): Pull in new filings, market data, sector / regulatory updates.
    - Triggered analyses when major events: earnings release, regulatory announcement, large price move, etc.
    - Peer benchmarking and scenario generation.
3. **User / Stakeholder Interface**
    - Analysts can view raw data, metrics, tweak assumptions).
    - Reports accessible by management / clients.
4. **Regulatory / Compliance Considerations (India / SEBI)**
    - Use only public / legally disclosed information.
    - Be cautious about forward-looking statements; label assumptions.
    - For PMS / advisory services ensure disclosures of risk, past returns, disclaimers.
    - Data privacy / corporate governance must be respected.

# 5. Expected Result and Progress

## 5.1 News Browsing Module

The news module gives users a clean, easy-to-read feed of financial stories. It pulls news from different sources but removes duplicates so the same story only shows once. Each story is shown as a simple card with a title, short summary, source, and finance tags. Users can filter or sort by time, source, topic, or their watchlist stocks. They can also see why an article appears, check related stories, save or share items, and search for similar content.

In the background, the system makes sure every article is well-formatted, tagged, and searchable, while keeping the original source for transparency. The interface is fast, supports accessibility, and offers basic personalization. A light cache keeps loading smooth, and simple APIs make it easy to connect with other services.

## 5.2 Stock Trend Module

### 5.2.1 Expected Result for Recommendation System & Prediction System

For the Recommendation subsystem within the Stock Trend Module, the following concrete outcomes are expected upon successful implementation:

At the initial stage, the Stock Trend Module will deliver a working pipeline that combines recommendation and prediction. The recommendation function will generate personalized stock lists that avoid simple popularity bias, offering varied results across sectors and

themes, each with a short explanation of why it is suggested. The prediction function will provide short-horizon (7–14 day) directional signals for individual stocks, returned quickly enough to be integrated into the user interface. Forecasts will include not only an up/neutral/down label but also a basic confidence score and a few main drivers for transparency. Together, these early capabilities will allow users to see more relevant stocks, explore multiple themes, and gain simple yet actionable insights into near-term movements

#### 5.2.2 Expected Result for RAG Q&A Module in Stock Trend module

In the initial phase, this module enables users to ask natural language questions about company filings, earnings calls, and recent news, and efficiently retrieves relevant document chunks from the past 90 days using vector search with ticker and time filters. Candidate results are reranked by similarity and recency, then passed to Qwen2.5-32B to generate evidence-based answers with source citations for transparency. The exposed /qa endpoint returns the answer, citations, retrieval stats, and a confidence score, while Redis caching and lightweight monitoring ensure stable and reliable responses within three seconds.

### 5.3 Combined Investment Recommendation Module

In its initial phase, the module is expected to improve analytic efficiency by reducing the time analysts spend on data collection and cleaning, allowing them to focus more on interpretation. It will produce reports with consistent metrics and transparent assumptions, making peer comparisons easier. Users will be able to respond more quickly to earnings releases, regulatory changes, and macro events, while also gaining better awareness of downside risks through scenario analysis and sensitivity checks. Over time, the system will support benchmarking of its outputs against actual market outcomes, providing a feedback loop on accuracy. For example, for a mid-cap stock, the module should be able to generate valuation estimates reasonably close to market consensus, even within a ±10–20% range, acknowledging market volatility.

## 6. Expected Challenges

### 6.1 News Browsing Module

**Large-scale copying** with minor modifications can cause content inflation and introduce ranking bias.

**Company aliases and cross-language naming** can lead to errors in linking entities to securities.

**Controlling system complexity and operational stability** while meeting low latency and high freshness requirements is challenging.

### 6.2 Stock Trend Module

### 6.2.1 Expected Challenges for Recommendation System

The development and deployment of the recommendation engine present several anticipated challenges that the project will need to address:

**Cold Start Problem:** A significant challenge will be providing accurate and engaging recommendations for new users who have no browsing history, making their profile vector non-existent or very sparse.

**Defining and Evaluating 'Success'**: While offline metrics like Recall@K are valuable, the true measure of success is user engagement and investment decisions, which are more difficult to quantify. How to accurately evaluate the recommendation engine's business impact will be a non-trivial challenge.

**Dynamic Nature of Financial Data:** User interests and market conditions change rapidly. A user's profile vector must be updated frequently to reflect their evolving preferences. How to determine the optimal update frequency and decay rate for the user profile to remain responsive without becoming overly volatile

### 6.2.2 Expected Challenges for Prediction System

**Market Noise and Non-Stationarity:** A core challenge will be designing models and retraining strategies that remain robust across regimes without overfitting to recent patterns.

**Balancing Short-Term vs. Long-Term Horizons:** Determining the optimal forecasting horizon and tuning models to maximize both statistical accuracy and user utility will be a persistent challenge.

**Feature Stability and Data Latency:** Ensuring stable feature engineering pipelines, timely ingestion, and monitoring for data drift will be critical to maintain reliable predictions.

### 6.2.3 Expected Challenges for RAG Q&A Module

**Data Quality & Preprocessing:** Financial reports, news, and conference call transcripts have complex formats and are prone to redundancy, noise, or omissions.Text segmentation strategies directly impact search performance and require continuous optimization.

**Retrieval Accuracy:** Vector search may recall fragments that are superficially similar to the question but semantically irrelevant. Over-reliance on semantic similarity may overlook timeliness and company relevance. Top-k and rerank strategies require a balance between recall and precision.

**Answer Faithfulness & Hallucination**：Even with search support, large models may still fabricate data or exaggerate conclusions.

### 6.3 Combined Investment Recommendation Module

| Challenge | Reason / Risk | Possible Mitigation |
|---|---|---|
| Model risk / over-fitting | Quant models might do well historically but perform poorly going forward; market regime changes (e.g. macro shocks) can invalidate assumptions. | Use out-of-sample testing; stress tests; include scenario analyses; build in model monitoring; update models periodically. |
| Interpretability & human trust | Users (analysts, clients) may not trust "black box" outputs, especially from AI/NLP; explanations are needed. | Make every recommendation accompanied by evidence, footnotes; allow drill-downs; include sensitivity tables; keep some human-in-loop. |
| Bias, noise, and errors from news / sentiment | False information, hype, errors; sentiment models may misinterpret context; overreaction to non-material events. | Use reliable sources; filter; validate events; have thresholds for triggering; allow human review for material items. |
| Maintaining freshness & up-dates | If the model becomes stale (e.g. macro environment shifts), or data feed breaks, output will degrade. | Set up monitoring; periodic retraining or recalibration; modular design so parts can be updated. |

## 7. Conclusion

*This implementation draws directly on concepts from our Intelligent Reasoning Systems course. By combining knowledge representation, reasoning under uncertainty, and data-driven inference, the system goes beyond raw computation to provide structured, explainable, and evidence-backed financial recommendations. The ability to integrate symbolic reasoning (rules, scenarios, constraints) with statistical methods (LLMs, embeddings, sentiment models) demonstrates how intelligent reasoning frameworks can be applied to real-world domains like financial analysis and investment decision-support.*