GenAI Powered Unstructured Data Transformation

Innovating Financial Investment Analysis and Decision Making

Company Advisor: **Jared Haite**
Faculty Advisor: **James Butler**


Lucy Liu


Sanya Chauhan

Business Context

Liberty Mutual Investments manages investments for Liberty Mutual

Current process utilizes earnings calls transcripts
1 analyst spends **> 100 hours/quarter** on reading an average of 100 transcripts

Project Objectives

Utilize **Large Language Models** to extract key insights

Identify and engineer novel **non-financial** features

Derive **structured time-series data**

Enable **quicker and enhanced** investment decisions

Data Scope

Used public earnings call transcripts obtained through web-scraping from investor relations sections of company websites.

1 Industry	3 Sectors	13 Companies	250 Transcripts	3 Million Words

Additional Data for Predictive Modeling:
Stock Prices, Trading Volume, and Company Revenue Data from Federal Reserve Economic Data (FRED), Yahoo Finance, and S&P Global.

Business Imapct

Saving 100+ hours

for 1 Analyst per Quarter by automtation

Increased productivity and manual efficiency

- Transitioned the process of extracting financial insight from fully manual efforts to automation utilizing LLMs in Earnings Call transcripts analysis
- Improved 91% of working efficiency for investment analysts, allowing them to focus on higher-level strategic activities

Strategic Insights

Innovative approaches motivated by GenAI

Unlocked Insights for Strategic Generative AI Use Case

- Inspired new applications in investment analysis to support the investment team and decision-making
- Uncovered novel features carrying valuable financial insights
- Provided a deeper understanding of market trends that were previously overlooked

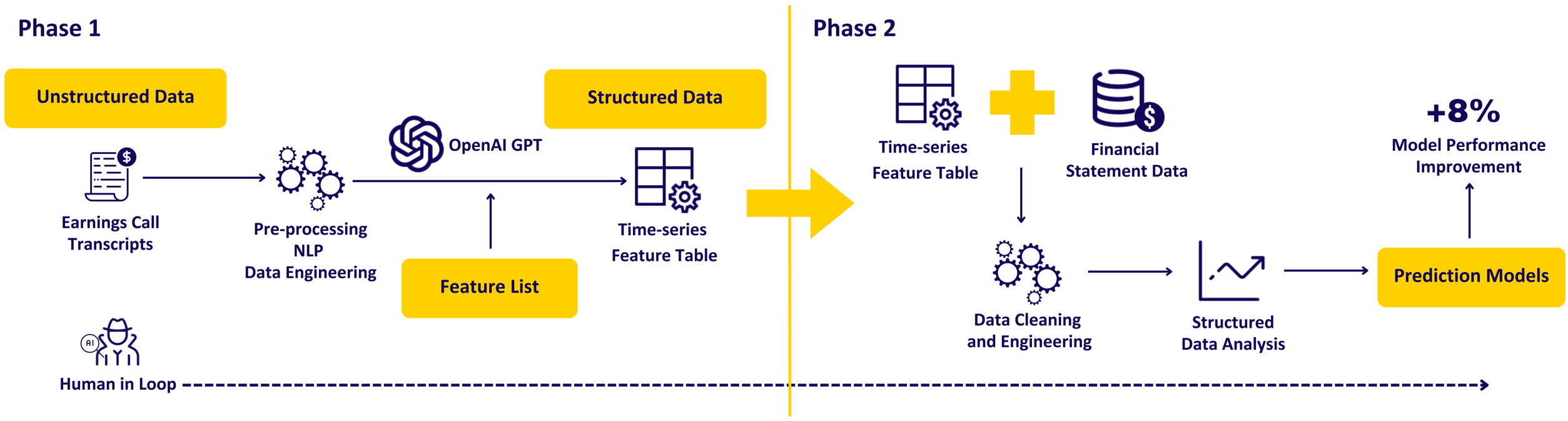
Novel Capability

for unstructured to structured data transformation

Developed Novel NLP Pipeline to Derive Structured Data

- Converted Earnings Call Transcripts into time-series data, enabling further predictive analysis over time
- Improved statistical modelings, resulting in a significant lift in prediction performance and providing more reliable forecasts

Methodology



1.1 Defining Features to Extract

Explored 4 avenues to define a list of non-financial features for data extraction

Independent Market Research → Manual Review of Sample Transcript → Expert Interviews with Liberty Mutual Investment → OpenAI GPT Model with Iterative Prompt Engineering

Focus on non-financial and forward looking features --> Defined a feature list

Observation: Different Terminologies for Same Context per Sector → **Solution:** Defining a Features' Synonyms Data Dictionary

Capturing holistic terminologies to ensure completeness

1.2 Feature Synonyms Dictionary

Feature	Manufacturing Sector	Services Sector	Transportation & Public Utilities Sector	Unique Terms across all Sectors
Employee Productivity	Human Operations Productivity, Labour Productivity	Human Capital Productivity, Employee Performance	Workforce Productivity	Human Operations Productivity, Labour Productivity, Human Capital Productivity, Employee Performance, Workforce Productivity
Stock-based Compensation	Stock Grants, Stock Options, Stock Compensation	Equity Compensation, Stock Options, Stock Compensation	Shared-based Compensation, Stock Compensation	Stock Grants, Stock Options, Stock Compensation, Equity Compensation, Shared-based Compensation

Sample Feature Synonyms Dictionary Generated using GPT Modeling

Note: Dotted box highlights the feature for which the data extraction example is shown in Section 2

2. Phase 1 - Structured Data Extraction

Deep Dive into Phase 1 Results

Table: Snipped of Cleaned Phase 1 Dataset

Company	Year	Quarter	Employee Productivity (in %)	Context for Employee Productivity	Tax Reform Impact	Inventory Satisfaction Expectations
Company 1	2023	Q2	-2	**Our workforce productivity was down 2% year-over-year** as daily car miles declined 6% in the quarter.	Positive	Improving
Company 2	2023	Q2	4	**Workforce productivity increased 4% year-over-year** despite the volume decrease.	Neutral	Stable

KEY TAKEAWAYS

Saves Time
Makes process of getting insights from Earnings Calls more efficient

Highlights Non-Financial Features
That might be traditionally overlooked

Enables Advanced Analysis
Data transformation pipeline allows data-driven decisions

Note: The complete extraction involved quarterly data 13 companies over 5 years.

3. Phase 2 - Predictive Modeling

Objective:
Test the impact of data extracted in Phase 1 on model performance by predicting Quarterly Close Stock Price.

Key Takeaway:
Improved model performance signifies importance and value in Phase 1 Data.

Baseline Model
Using traditional financial features

Current Model
Using additional Phase 1 extracted data

Best Model
Ridge regression: Robustness against missing values and overfitting

MODEL PERFORMANCE

Measured by R-Squared

Baseline: **0.88**

Current Model: **0.96**