


# GPT vs Analyst EPS Predictions

Exploring Prediction Accuracy, Errors,  
and Financial Implications

By: Palvi Sharma & Kartik Joshi

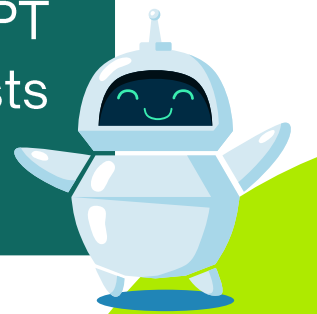


# MOTIVATION



The increasing adoption of AI and large language models (LLMs) like ChatGPT in financial decision-making raises critical questions about their predictive power and reliability. Financial analysts rely on deep market insights, experience, and macroeconomic factors to make earnings per share (EPS) predictions, but **can a publicly available AI model like ChatGPT match or even outperform them?**

This study explores whether ChatGPT's EPS predictions are as accurate as those of professional analysts, assessing its error rates, consistency, and correlation with actual earnings. If proven reliable, this could have massive implications for investors, businesses, and financial institutions, potentially reshaping how forecasting is done. However, if ChatGPT struggles in certain scenarios, it raises concerns about the risks of AI-driven financial decision-making. Our analysis provides a comprehensive evaluation to determine whether ChatGPT can be a trustworthy tool for financial predictions or if human analysts still hold a decisive edge.



# Financial Statement Analysis with Large Language Models (Kim, Muhn, & Nikolaev, 2024)

This study asks whether large language models (LLMs) – specifically GPT-4 – can perform financial statement analysis similarly to a professional analyst. The core question is if an LLM can analyze a firm's financial statements to predict the direction of future earnings changes (increase or decrease) and how its performance compares to human financial analysts and traditional models.

## METHODOLOGY COMPARISON

### Used in the Research Paper

- The researchers used a comprehensive dataset of Compustat annual financial statements (1968–2021)
- GPT-4 (through the GPT-4 Turbo API) was given the numeric balance sheet and income statement data for a firm and instructed to analyze them.

### Innovated by us for this project

- We used the IBES dataset to calculate the EPS of US Companies (2020–2024)
- Chat GPT-4o (public model) was given the numeric balance sheet and income statement data of top 40 S&P 500 companies.



# Dataset Overview

For this study, we constructed a dataset comprising earnings per share (EPS) forecasts. The goal was to evaluate the accuracy, error distribution, and reliability of GPT-based forecasts compared to traditional analyst estimates.

## Data Source:

- Financials & EPS predictions for 40 S&P 100 companies (2020-2024) from WRDS.
- Data from Yahoo Finance API for GPT Predictions

## Key Variables:

- Actual EPS (reported by companies).
- Forecasted EPS (Analyst estimates).
- GPT Predicted EPS (LLM-generated predictions).

## Key Metrics:

Mean Absolute Error (MAE), Accuracy Trends, Forecast Errors.

The CONTENTS Procedure			
Data Set Name	IBES.DET_EPSUS	Observations	33720359
Member Type	DATA	Variables	27
Engine	V9	Indexes	9
Created	12/12/2024 16:40:00	Observation Length	184
Last Modified	12/12/2024 16:41:00	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label	I/B/E/S Detail History - Detail File with Actuals (EPS for US Region)		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	94989
First Data Page	1
Max Obs per Page	355
Obs in First Data Page	323
Index File Page Size	8192
Number of Index File Pages	489829
Number of Data Set Repairs	0
Filename	/wrds/ibes/sasdata/det_epsus.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	2485598143
Access Permission	rw-r-----

### Income Statement:

	Tax Effect Of Unusual Items	Tax Rate For Calcs	Normalized EBITDA
2024-09-30	0.0	0.241	134661000000.0
2023-09-30	0.0	0.147	125820000000.0
2022-09-30	0.0	0.162	130541000000.0
2021-09-30	0.0	0.133	123136000000.0
2020-09-30	NaN	NaN	NaN

	Net Income From Continuing Operation	Net Minority Interest	\
2024-09-30	93736000000.0		
2023-09-30	96995000000.0		
2022-09-30	99803000000.0		
2021-09-30	94680000000.0		
2020-09-30	NaN		

	Reconciled Depreciation	Reconciled Cost Of Revenue	EBITDA
2024-09-30	11445000000.0	210352000000.0	134661000000.0
2023-09-30	11519000000.0	214137000000.0	125820000000.0
2022-09-30	11104000000.0	223546000000.0	130541000000.0
2021-09-30	11284000000.0	212981000000.0	123136000000.0
2020-09-30	NaN	NaN	NaN

	EBIT	Net Interest Income	Interest Expense	...	\
2024-09-30	123216000000.0	NaN	NaN	...	
2023-09-30	114301000000.0	-183000000.0	3933000000.0	...	
2022-09-30	119437000000.0	-106000000.0	2931000000.0	...	
2021-09-30	111852000000.0	198000000.0	2645000000.0	...	
2020-09-30	NaN	890000000.0	2873000000.0	...	



# TEST PROCEDURE



## Data Collection & Preprocessing

- Gathered Actual EPS data from financial reports of 40 S&P 100 companies (2020–2024).
- Collected Analyst EPS forecasts from market databases.
- Generated GPT EPS predictions using ChatGPT’s publicly available version, ensuring it had access only to general financial trends and not proprietary or insider data. Standardized all EPS values for consistency across sources.



## Error Calculation

We evaluated the performance of both GPT and Analyst forecasts by calculating their prediction errors relative to actual EPS:

- Mean Absolute Error (MAE): Measures average prediction error magnitude.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Predicted\_EPS_i - Actual\_EPS_i|$$

- Absolute Error Distribution: Analyzed the spread and variance of errors for both models.



## Comparative Analysis

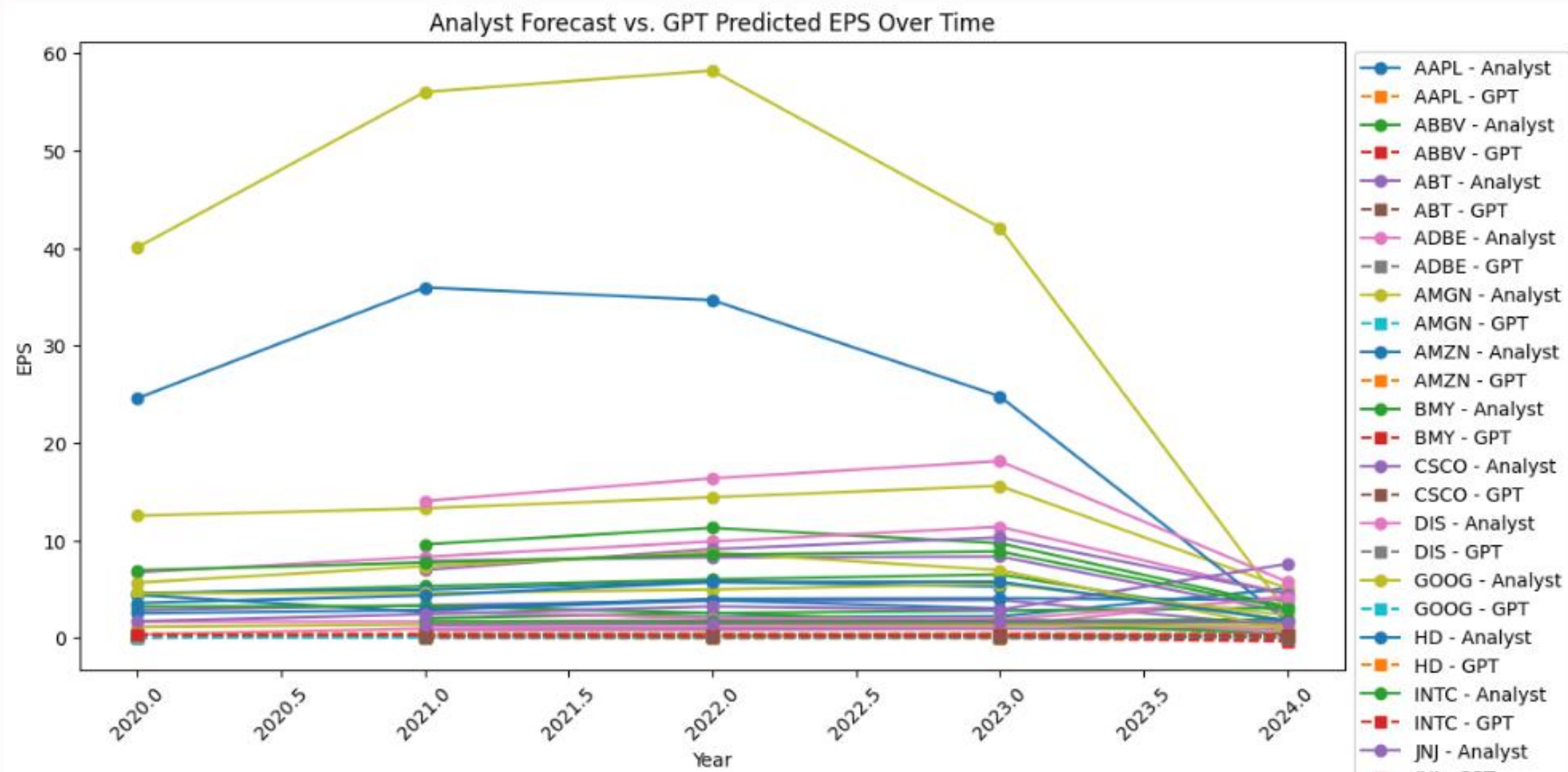
- Direct Prediction Comparison: Plotted GPT vs Analyst forecasts to observe systematic biases.
- Trend Over Time: Measured prediction accuracy trends (2020–2024) to check if GPT improved or worsened over time.
- Correlation Analysis: Evaluated whether GPT’s predictions correlated more or less with actual EPS than analysts’ forecasts using a heatmap.



## Statistical Significance Testing

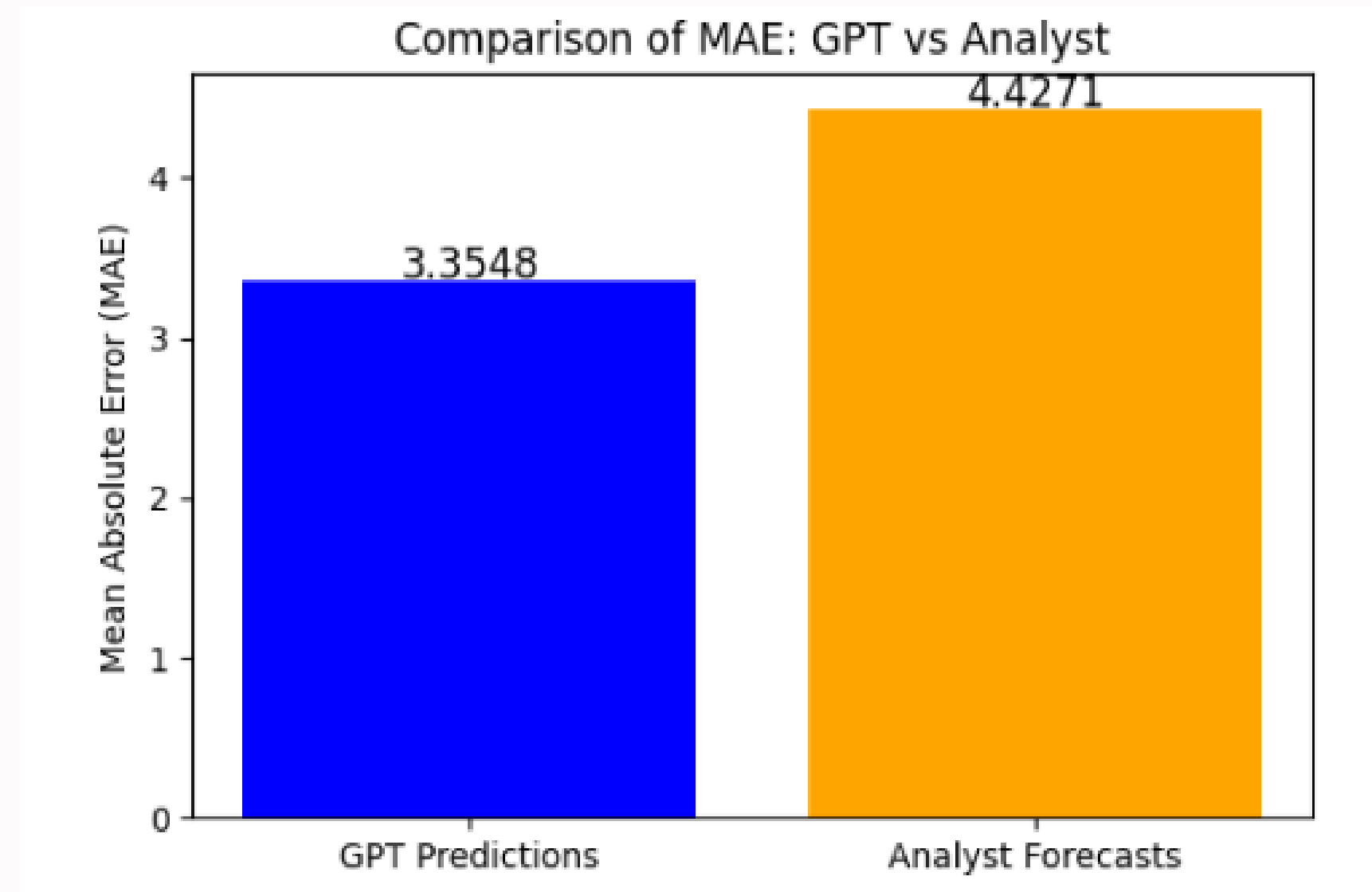
- Paired T-Test: Tested if the mean difference in errors between GPT and Analysts was statistically significant.
- Wilcoxon Signed-Rank Test: A non-parametric test to verify if GPT systematically outperforms analysts.

# Analyst vs GPT Predictions Over Time



- Some companies show consistent EPS growth trends, while others have fluctuations.
- GPT and analysts track each other in some cases, but GPT predictions are consistently more conservative.
- GPT follows the general trend but is less extreme in high EPS cases.

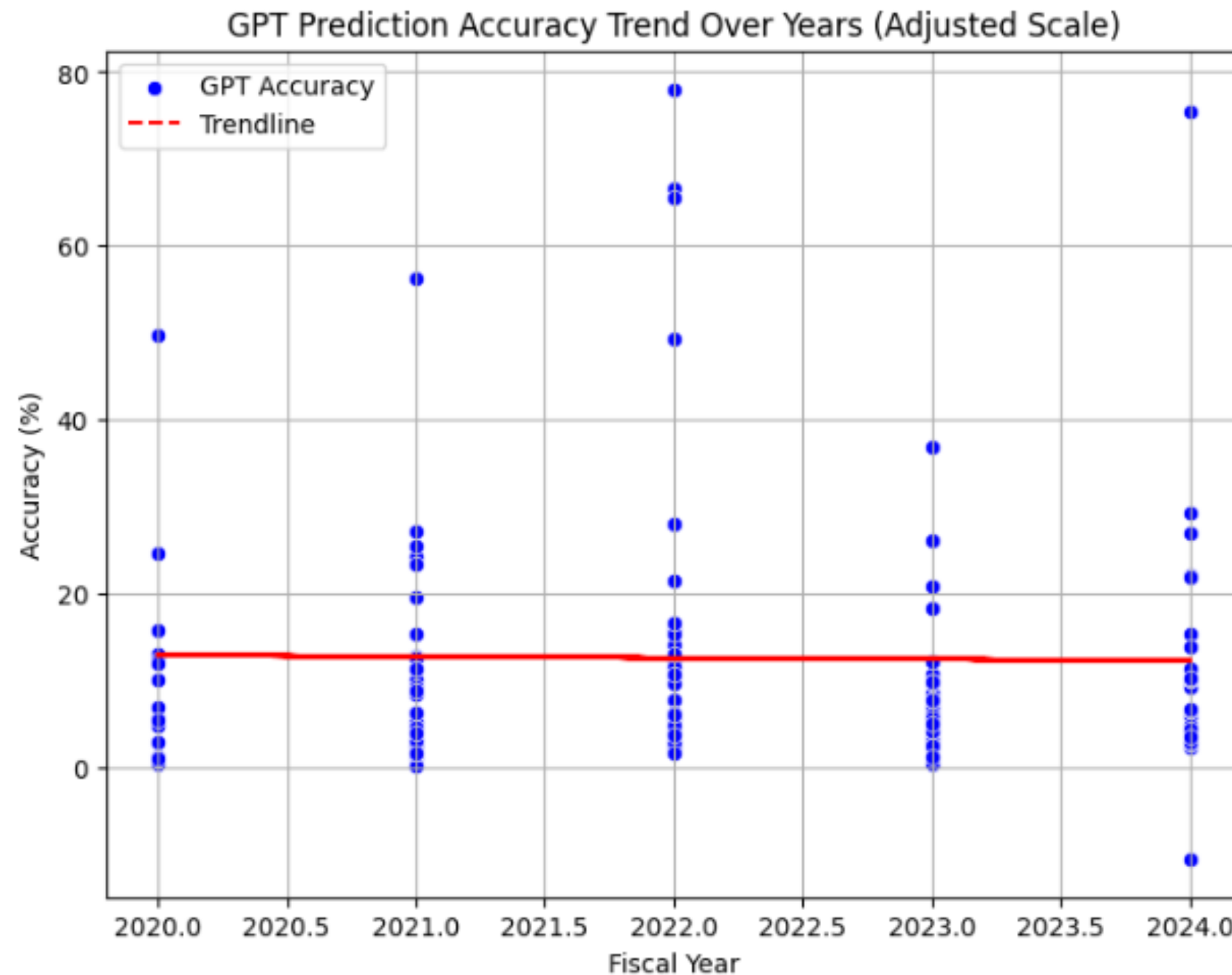
# MAE (Mean Absolute Error) Comparison



GPT MAE = **3.35**, Analyst MAE = **4.42**

GPT has a lower MAE, meaning it is closer to actual EPS than analysts on average. GPT beats analysts in MAE, meaning its average prediction is more accurate. Analysts have higher variance, leading to higher mean absolute errors.

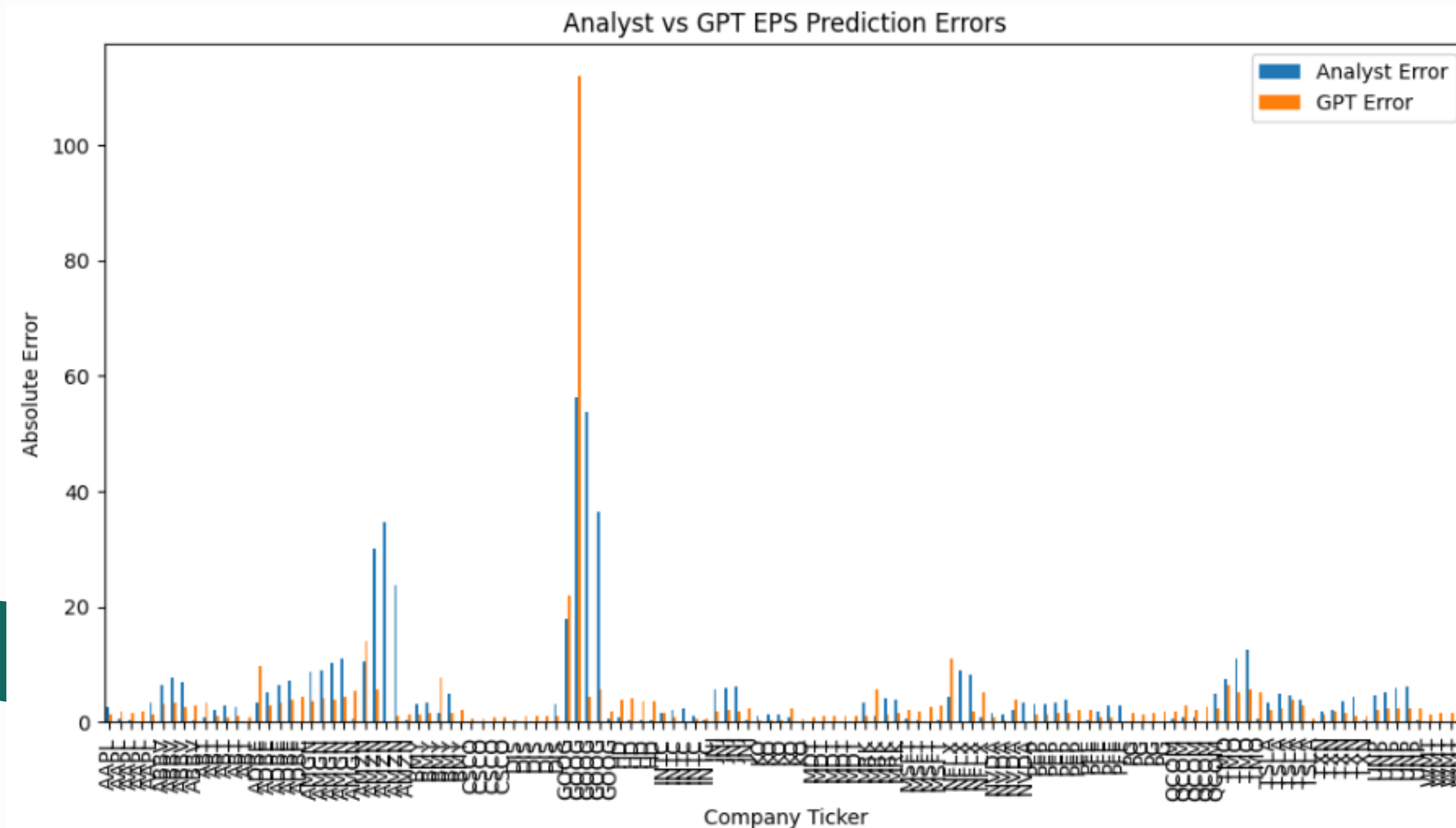
# GPT Accuracy Over Time (Trendline)



- GPT accuracy remains stable over the years. Some outliers exist, but the trendline suggests consistency.
- GPT has not significantly improved or declined in accuracy over time.
- External factors (macroeconomic shifts) may not impact GPT predictions as much as analysts.

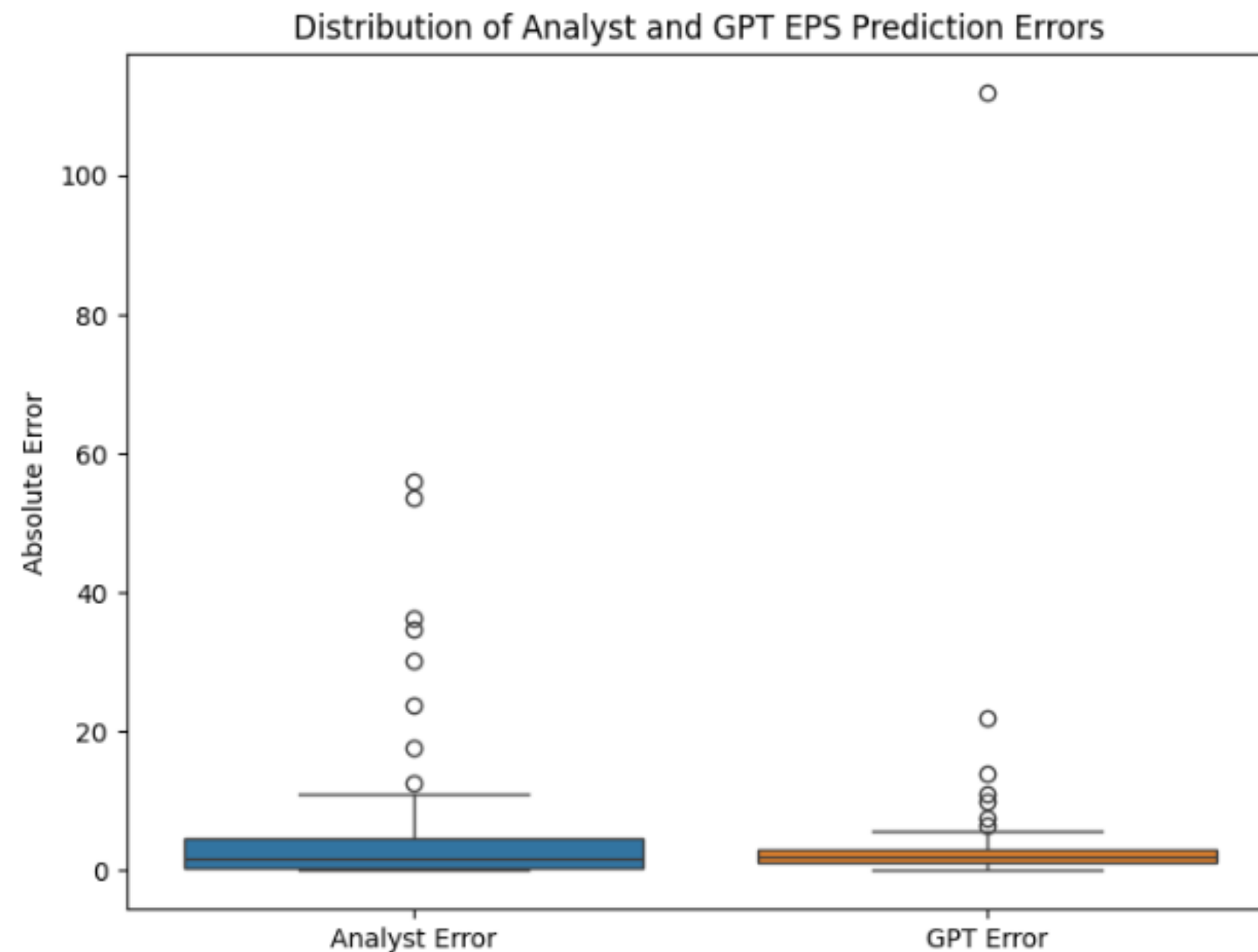


# Analyst vs GPT (EPS Prediction Errors Chart)



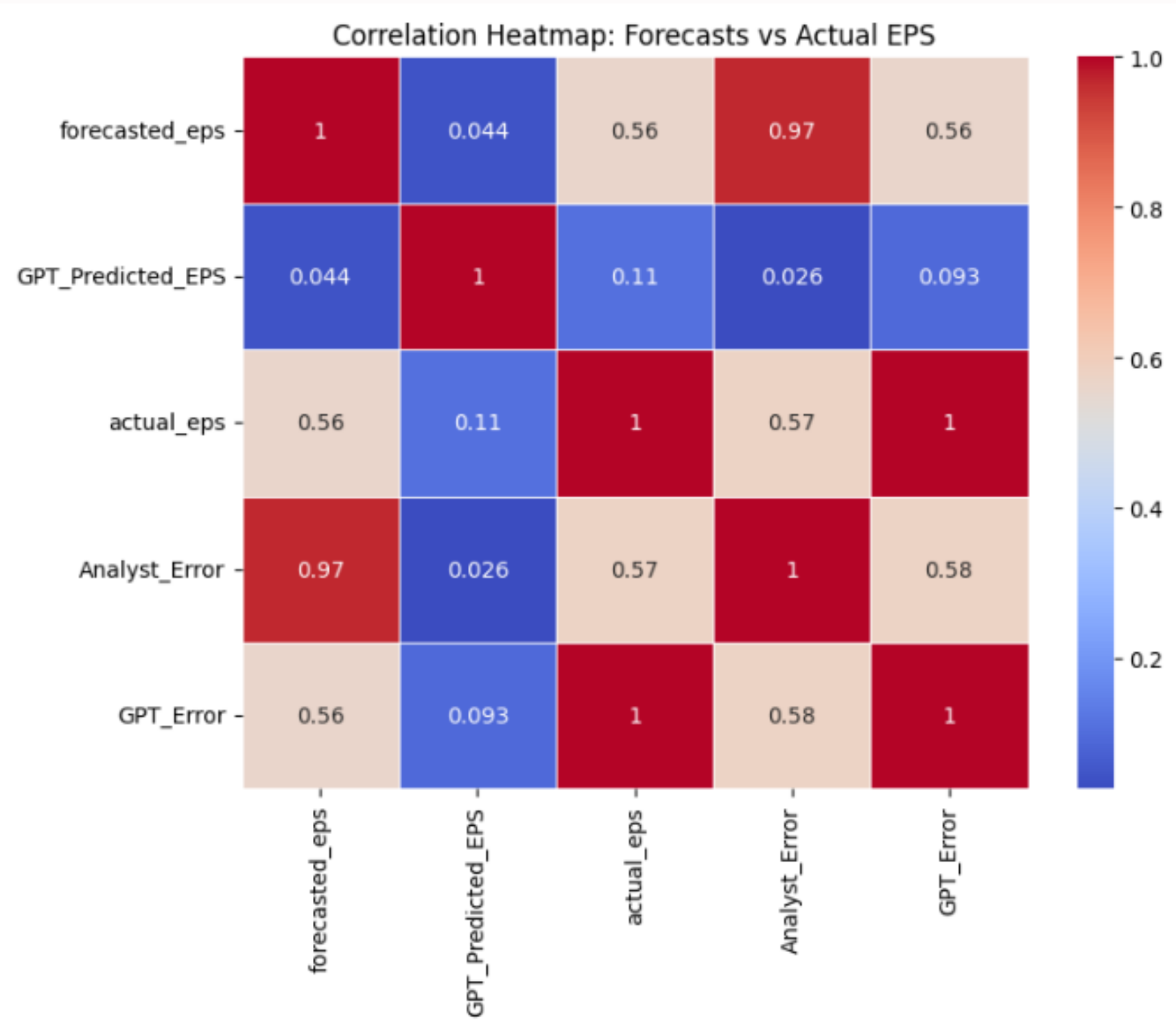
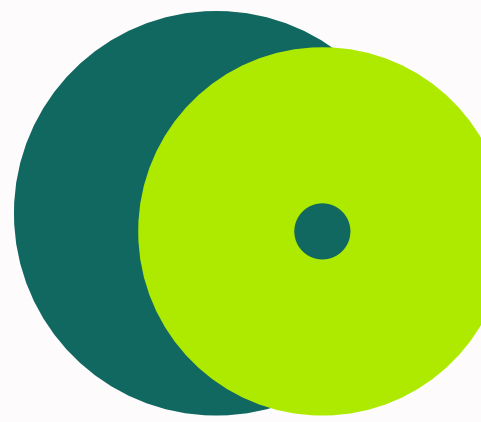
- In many cases, analyst errors (blue bars) are larger than GPT errors (orange bars). However, there are instances where GPT has significantly higher errors, meaning it sometimes makes more extreme mistakes than analysts.
- GPT often outperforms analysts but has large errors in certain cases. Extremely high GPT errors indicate it struggles with certain firms, possibly due to volatility or data limitations.

# Distribution of Analyst vs GPT Prediction Errors



- Median error (central line) is lower for GPT than analysts. Analysts have higher variance and more extreme outliers.
- GPT errors are lower on average, but it still makes some extreme mistakes. Its errors are more tightly distributed, suggesting more consistency.

# Correlation Heatmap: EPS Forecasts vs Actual EPS



- Analyst EPS is more correlated with actual EPS (0.56) than GPT (0.11).
- GPT errors are less correlated with actual EPS, meaning it may struggle with certain types of data.
- GPT introduces more variance and does not always follow analysts. It may struggle with industry-specific or macroeconomic influences that analysts consider.

# FINAL TAKEAWAYS

GPT shows promise in financial forecasting, with lower mean errors and higher consistency than analysts. GPT shows promise in financial forecasting, with lower mean errors and higher consistency than analysts.

**GPT as a Tool,  
Not a  
Replacement**

GPT predictions are data-driven, meaning they may be too reliant on past trends and fail to adjust for sudden market shifts. This is a key limitation in financial forecasting, where forward-looking analysis is often more valuable than historical patterns.

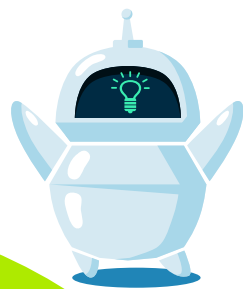
**Overfitting to  
Historical Data**

GPT may perform well in stable industries but struggle in sectors with rapid innovation (e.g., tech, biotech). Further segmentation by sector and market conditions is needed to determine where GPT is most reliable.

**Industry-Specific  
Performance  
Varies**

Unlike analysts, GPT lacks access to real-time earnings calls, company guidance, and macroeconomic shifts. Analysts incorporate CEO statements, Federal Reserve policies, geopolitical risks, and other non-quantitative signals that GPT does not process effectively.

**Lack of Market  
Context & Forward-  
Looking Factors**



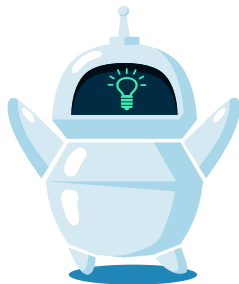
## Investment Implications

### Can Investors Rely on GPT for Financial Forecasting?

- GPT can serve as a complementary forecasting tool, but it should not replace human analysts entirely.
- Investors seeking consistent, low-variance predictions may find GPT valuable, especially for benchmarking multiple forecasts.
- However, GPT lacks the qualitative judgment and forward-looking market analysis that analysts incorporate, making it less suited for high-growth or volatile sectors.



# FINAL TAKEWAYS



## Risk Management & Portfolio Strategies

- GPT's lower variance and fewer extreme errors make it useful for risk-averse investment strategies, where avoiding large prediction failures is a priority.
- However, GPT's inability to make aggressive high-EPS predictions suggests that it may miss out on high-growth opportunities, which could be a drawback for active investors or hedge funds.

## Future Improvements & Potential Applications

- GPT's performance could improve with real-time financial news integration and access to economic indicators, earnings calls, and sector-specific data.
- Future AI models trained specifically on financial data may bridge the gap between AI and human analysts, potentially making AI-driven forecasting a major disruptor in financial markets.

CONTRIBUTIONS TABLE

NAME	%age	Tasks
Kartik	50	Ideation; Prediction Implementation; API Extraction
Palvi	50	Data Extraction & Filtration; Interpretation & Presentation



Thank You