

Predicting the Stock Market

Manav Patel

Question:

Can historical trading data and moving averages improve LSTM model accuracy in predicting future stock prices of technology and pharmaceutical companies?

Background:

The stock market enables individuals and institutions to buy and sell ownership shares of publicly traded companies. Shareholders may benefit from capital appreciation as company value increases and, in some cases, dividend distributions. Stock prices are determined by supply and demand dynamics and are influenced by company-specific fundamentals, broader economic conditions, and market sentiment (Akana, Drayton, and Lee).

Daily stock price data typically include the opening price, highest price, lowest price, and closing price, along with trading volume. These variables provide a structured summary of intraday market behavior and form the basis for most quantitative financial analyses. Over time, historical price and volume data are commonly used to identify trends, volatility, and momentum.

One widely used technique for trend analysis is the moving average. Simple moving averages (SMAs) compute the mean price over a fixed window, while exponential moving averages (EMAs) place greater weight on recent observations (Monfared). Moving averages help smooth short-term fluctuations and are frequently used as technical indicators in financial modeling.

Although historical price patterns can reveal trends, stock markets are influenced by numerous exogenous factors including macroeconomic conditions, geopolitical events, and firm-specific news, which limit predictability (Beers). As a result, price forecasting models must be evaluated cautiously and with appropriate performance metrics.

Daily stock trading data include the opening, highest, lowest, and closing prices for each trading day. It also includes trading volume, which measures the number of shares traded. Historical price data may also be affected by corporate actions such as dividends and stock splits (Beers). Dividends are distributions of a company's profits to shareholders, whereas stock splits increase the number of outstanding shares without changing the company's overall market capitalization (Akana, Drayton, and Lee).

Moving averages are widely used technical indicators that reduce the impact of short-term price volatility. Simple moving averages (SMAs) calculate the average price over a specified period, while exponential moving averages (EMAs) have greater emphasis on more recent prices. These indicators are commonly applied to identify trends and momentum in financial time series data.

These key terms are important because it is necessary for understanding stock analysis. Historical trading data, which includes these metrics over time, is used for understanding patterns and trends of an underlying asset (eg. S&P 500, NASDAQ, etc.). An index is a

managed portfolio of several companies. For instance, a rising moving average might suggest an upward trend meaning that it could be a good time to buy. However, a falling trend could be a good time to sell. Although predictive models can be useful for identifying trends, real-world accuracy remains limited due to external economic and behavioral factors.

The stock market reflects broader economic health and global wealth distribution. In 2023, the global stock market capitalization went past \$100 trillion which shows the scale of it. Predicting future prices helps investors hedge against risks and maximizes return which all in the end makes them more money. For example, during the COVID-19 pandemic, pharmaceutical stocks went up due to vaccine development, while technology companies' stocks benefited from remote working (Akana, Drayton, and Lee). Accurate predictions can prevent losses and help people gain a lot of money. Many factors play a role in the prices of stocks, some are: random affairs going on in the world, natural disasters, news, inflation, company performance, supply and demand, etc. In other words, stock movements are highly stochastic and influenced by many external variables, which limits its predictability (Beers).

In the Tech sector, companies like Apple(AAPL), Google(GOOG), and NVIDIA(NVDA) are driven by innovation, consumer demand, and technological breakthroughs. Tech stocks are often volatile, with rapid growth potential but also sharp declines from competition or changes (Velasquez). The Pharmaceutical sector, like AbbVie (ABBV), Eli Lilly (LLY) and Pfizer (PFE), focuses on drug development, approvals, and patents. Pharma stocks tend to be more stable, offering dividends, but are sensitive to clinical trial outcomes, FDA decisions, and patent expirations (Speights). Analyzing historical data and moving averages in these sectors is important to predict the future prices because it reveals sector-specific patterns. Tech might show trends from AI advancement, while pharma might show stable cyclical behaviour based on its drug pipelines or have a breakthrough and surge. These specific companies' behavior is highly based on external factors and have a huge market capitalization (Beers). Real life events impact these industries and specific companies differently.

This research is important for many reasons—first, it provides investment knowledge, which helps retail investors who are part of about 25% of the U.S. trading volume using apps such as Robinhood, etc (Mikulic). Second, it contributes to financial literacy, and per the surveys, only 43% of Americans are stock market literate (Akana, Drayton, and Lee). Third, in an era of AI and ML with a lot of data, using machine learning on historical data to build AI will help in improving prediction accuracy beyond the original methods. Finally, comparing tech and pharma shows how different industries respond to economic factors.

Introduction:

Using LSTM models trained on historical data and moving averages, this study compares the predictive accuracy across the tech and pharmaceutical sectors. Specifically, six companies: three tech (Apple, Google, and Nvidia) and three pharma (AbbVie, Eli Lilly, and Pfizer). This paper is timely as this is an era of AI and ML and can be used for historical data and MA analysis to predict future prices, however the accuracy could be very inaccurate due to the many factors that play a role in the market.

The research utilizes historical data from each companies' listing date (for example: ABBV from 2013) to the current date sourced by yfinance. I created an AI model to extract data to visualize trends, correlations, and distributions and use it to predict future prices. Moving averages (10-day, 20-day, 50-day SMAs) have been calculated to capture short, medium, and long term trends, as they help filter noise in the data. I created machine learning models, particularly Long Short-Term Memory (LSTM) networks in TensorFlow and Pytorch, to predict the closing, open, low, and high prices for the next ten days from July 21 to July 30 to see the AI models' accuracy. LSTM networks are a type of recurrent neural network designed to model sequential data by maintaining an internal memory state (Sayah). This architecture allows LSTMs to capture temporal dependencies across multiple time steps while mitigating the vanishing gradient problem that affects standard recurrent networks (Monfared). So LSTM networks are ideal for sequential data like stocks, since they remember long-term dependencies while handling gradients that are not there (Sayah).

I used correlation heatmaps and distribution plots to identify relations among stock variables before training the models. I then train LSTM (Long short term memory) models: a TensorFlow model using Open, High, Low, Volume, and Close and PyTorch models incorporating SMAs. The dataset was split chronologically to preserve temporal order. The first 80% of observations were used for training, the subsequent 10% for validation, and the final 10% for testing with 50-100 epochs and MSE loss. This approach ensures that the model is evaluated on future data relative to its training period. Predictions were also made for 10 and 100 days starting on July 21, 2025 using realistic simulations incorporating historical volatility and sector specific events such as drug approvals for pharma and AI news for tech.

This study investigates whether incorporating historical trading data and moving averages improves the predictive performance of Long Short-Term Memory (LSTM) models for stock price forecasting. Specifically, the research evaluates LSTM-based time-series models across six publicly traded companies: three from the technology sector and three from the pharmaceutical sector.

Historical price data are widely available and frequently used in quantitative finance, making them a practical foundation for time-series forecasting models. LSTM networks are well suited for sequential data because they can model temporal dependencies while mitigating vanishing gradient issues common in standard recurrent neural networks.

The objective of this study is not to claim precise price prediction, but to evaluate whether adding moving average features improves short-term forecasting accuracy relative to models trained on raw price and volume data alone.

Methods:

The approach for this project involves retrieving historical stock data, conducting EDA (exploratory data analysis), computing movement, and applying LSTM models for predictions.

Data was downloaded using yfinance from 2010 to August 2025, including the Open, High Low, Close, Volume, Dividends, and Stock Splits. Exploratory data analysis was also included with

printing shapes, statistical summaries, unique values, correlation heatmaps, pairplots, violin and box plots with statistics, line plots for attributes (daily, weekly, monthly), predicted candlestick charts, yearly mean bars, histograms, SMA plots, and risk return scatter plots.

LSTM networks were utilized, which is a type of recurrent neural network made for time series data (Monfared). They were a good choice because stock prices show patterns over time and LSTMs are able to remember past values when making predictions. Other models like simple feedforward networks do not handle this kind of data as well. I also added simple moving averages as features because they smooth out random ups and downs and show the overall trend more clearly. This helps the model focus on momentum and direction instead of just noisy daily changes.

In TensorFlow, feature scaling was applied to improve numerical stability during neural network training, as LSTM models are sensitive to differences in feature magnitude. Min-Max scaling parameters were computed using only the training data and then applied to validation and test sets to prevent data leakage (Sayah). Because scaling parameters are derived from historical training data, values outside the training range may be clipped or extrapolated, which represents a limitation of this preprocessing approach. I then made sequences of 10 days of stock prices so that the model would predict the closing price on the 11th day. The dataset was split into 80 percent training, 10 percent validation, and 10 percent testing. The model had two LSTM layers with 64 and 32 units, followed by two dense layers with 25 and 1 units. I trained it with the Adam optimizer and mean squared error loss for 50 epochs.

In PyTorch, I included not only the open, high, low, close, and volume but also SMA features for 10 days, 20 days, or both. The data was scaled between zero and two and sequences were created in the same way. I used DataLoaders with a batch size of 32. The model had an LSTM layer with 64 hidden units, a dropout layer to reduce overfitting, and a linear output layer. I trained for 50 to 100 epochs with Adam and mean squared error loss and printed the training and validation losses every few epochs to keep track of progress.

For predictions, I made both short term and long term forecasts. I predicted the next 10 days of closing prices, along with open, high, low, and volume values that followed basic stock logic such as the low being less than or equal to the open and close, and the high being the maximum. I also created 100 day simulations starting from July 21, 2025. Long-horizon forecasts (100 days) are included solely for qualitative visualization and should not be interpreted as reliable predictions. Error accumulation in recursive time-series forecasting makes long-term predictions increasingly unstable, particularly in financial markets. These longer forecasts included randomness and sector specific events, like drug approvals in pharmaceuticals and new AI product launches in tech, to make them more realistic. Candlestick charts were also used to visualize 30 day periods and to compare real and predicted results.

This approach is appropriate because it combines stock trends with a machine learning model that is good at handling sequences. It lets us test how much moving averages improve the predictions and also compare the results between tech and pharmaceutical companies. It also lets us see how accurate the model is.

Data Collection

Historical daily stock data were retrieved using the yfinance Python library for the period January 1, 2010 through July 20, 2025. The dataset included open, high, low, close, and volume values. Dividend and stock split information were retrieved but not used as model inputs.

Feature Engineering

Simple moving averages (SMAs) with window sizes of 10, 20, and 50 trading days were computed using closing prices. These features were included to capture short-term and medium-term trends while reducing high-frequency noise.

Data Preprocessing and Scaling

All continuous input features were normalized using Min-Max scaling to improve numerical stability during neural network training. Scaling was performed using parameters computed only on the training set, which were then applied consistently to validation and test sets to prevent data leakage.

Scaling was necessary because neural networks are sensitive to feature magnitude; without normalization, features such as trading volume can dominate the loss function and impair convergence (Alberto, Einhorn, Fisch, Le, and Sautter).

Train-Validation-Test Split

The dataset was split chronologically to preserve temporal order:

Training set: first 80% of observations

Validation set: next 10%

Test set: final 10%

This temporal split ensures that the model is evaluated on future data relative to its training window, which more accurately reflects real-world forecasting conditions.

Sequence Construction

Input sequences consisted of rolling windows of 10 consecutive trading days, with the model predicting the closing price on day 11. This window size was chosen to balance short-term temporal context with computational efficiency.

Model Architecture

LSTM models were implemented in both TensorFlow and PyTorch. The TensorFlow model consisted of two stacked LSTM layers with 64 and 32 hidden units, followed by two fully

connected layers. The PyTorch model included a single LSTM layer with 64 hidden units, a dropout layer to reduce overfitting, and a linear output layer.

Training Procedure

Models were trained using the Adam optimizer and mean squared error (MSE) loss. Training was conducted for 50–100 epochs depending on convergence behavior, with validation loss monitored to detect overfitting.

Evaluation Metrics (Required)

Model performance was evaluated using:

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

These metrics were computed on the held-out test set and used to compare models trained with and without moving average features.

Discussion:

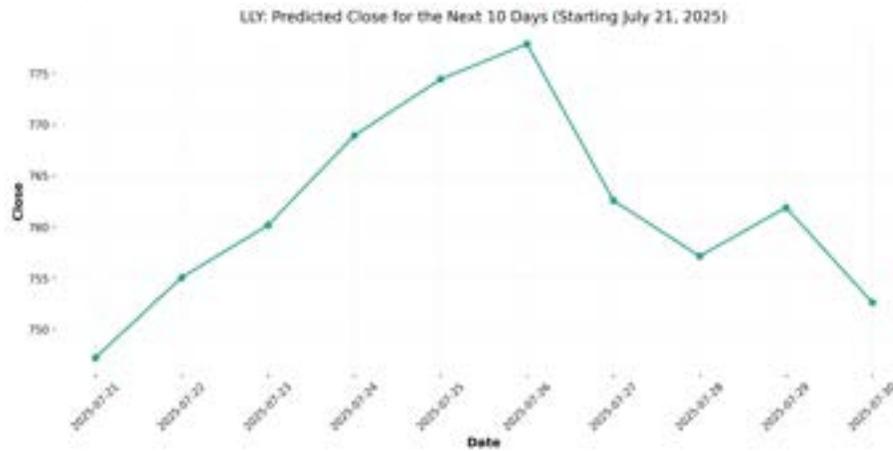
Running the closing price models created:

Abbvie:

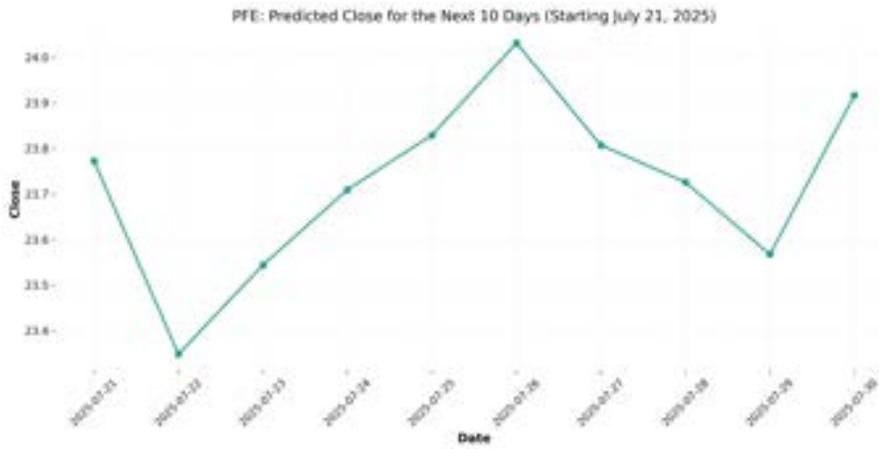




Eli Lilly:

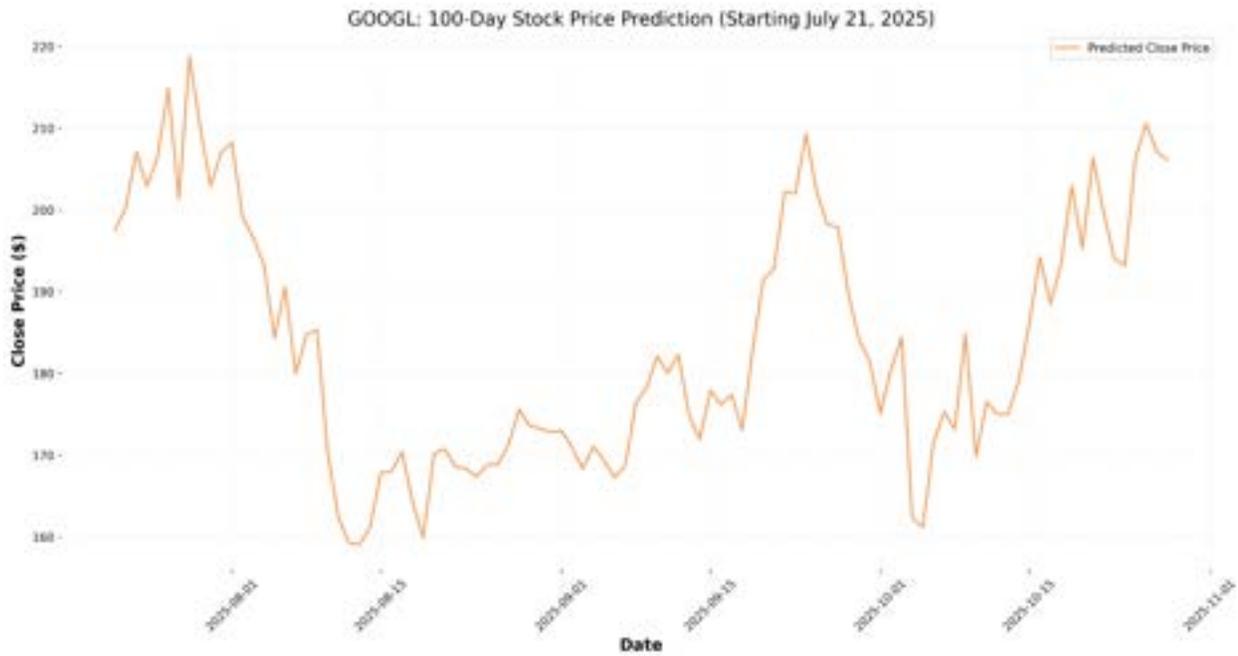


Pfizer:



Google:

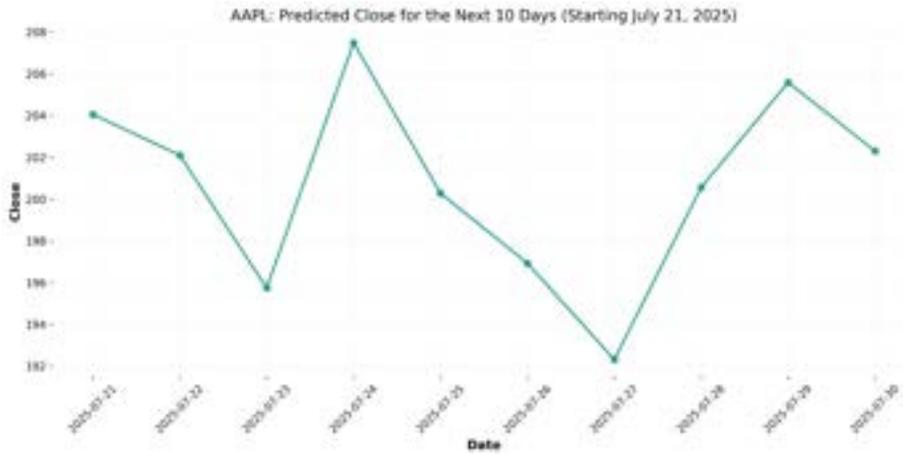




Nvidia:



Apple:



Prediction Accuracy Table:

| | Apple Predicted | Apple Actual | Google Predicted | Google Actual | Nvidia Predicted | Nvidia Actual | Abbvie Predicted | Abbvie Actual | ES Lilly Predicted | ES Lilly Actual | Pfizer Predicted | Pfizer Actual |
|-----------|-----------------|--------------|------------------|---------------|------------------|---------------|------------------|---------------|--------------------|-----------------|------------------|---------------|
| 7/21/2025 | \$204.06 | \$212.48 | \$184.20 | \$181.15 | \$186.50 | \$171.38 | \$189.00 | \$184.85 | \$747.00 | \$752.18 | \$23.77 | \$24.26 |
| 7/22/2025 | \$202.11 | \$214.40 | \$189.80 | \$192.11 | \$186.00 | \$167.03 | \$189.00 | \$187.11 | \$755.00 | \$776.44 | \$23.35 | \$25.14 |
| 7/23/2025 | \$186.75 | \$214.15 | \$190.90 | \$191.51 | \$190.00 | \$170.78 | \$182.00 | \$186.55 | \$760.00 | \$798.89 | \$23.55 | \$25.36 |
| 7/24/2025 | \$207.48 | \$213.70 | \$186.10 | \$183.20 | \$197.00 | \$173.74 | \$190.50 | \$190.83 | \$769.00 | \$805.43 | \$23.70 | \$25.35 |
| 7/25/2025 | \$200.29 | \$213.88 | \$184.30 | \$184.08 | \$194.00 | \$173.50 | \$190.00 | \$190.28 | \$774.00 | \$812.69 | \$23.85 | \$24.79 |
| 7/26/2025 | \$198.94 | \$214.05 | \$186.00 | \$183.42 | \$190.00 | \$178.75 | \$190.80 | \$188.52 | \$778.00 | \$808.11 | \$24.02 | \$24.31 |
| 7/27/2025 | \$182.32 | \$214.05 | \$184.30 | \$183.42 | \$191.50 | \$178.75 | \$191.80 | \$188.52 | \$782.00 | \$808.11 | \$23.80 | \$24.31 |
| 7/28/2025 | \$200.57 | \$214.05 | \$185.80 | \$185.43 | \$197.00 | \$178.75 | \$195.50 | \$188.52 | \$757.00 | \$808.11 | \$23.72 | \$24.31 |
| 7/29/2025 | \$205.59 | \$211.27 | \$186.80 | \$185.43 | \$211.00 | \$175.51 | \$186.80 | \$191.22 | \$782.00 | \$782.85 | \$23.57 | \$24.30 |
| 7/30/2025 | \$202.30 | \$209.05 | \$185.40 | \$187.44 | \$211.50 | \$179.27 | \$201.00 | \$189.31 | \$752.00 | \$780.08 | \$23.92 | \$23.81 |

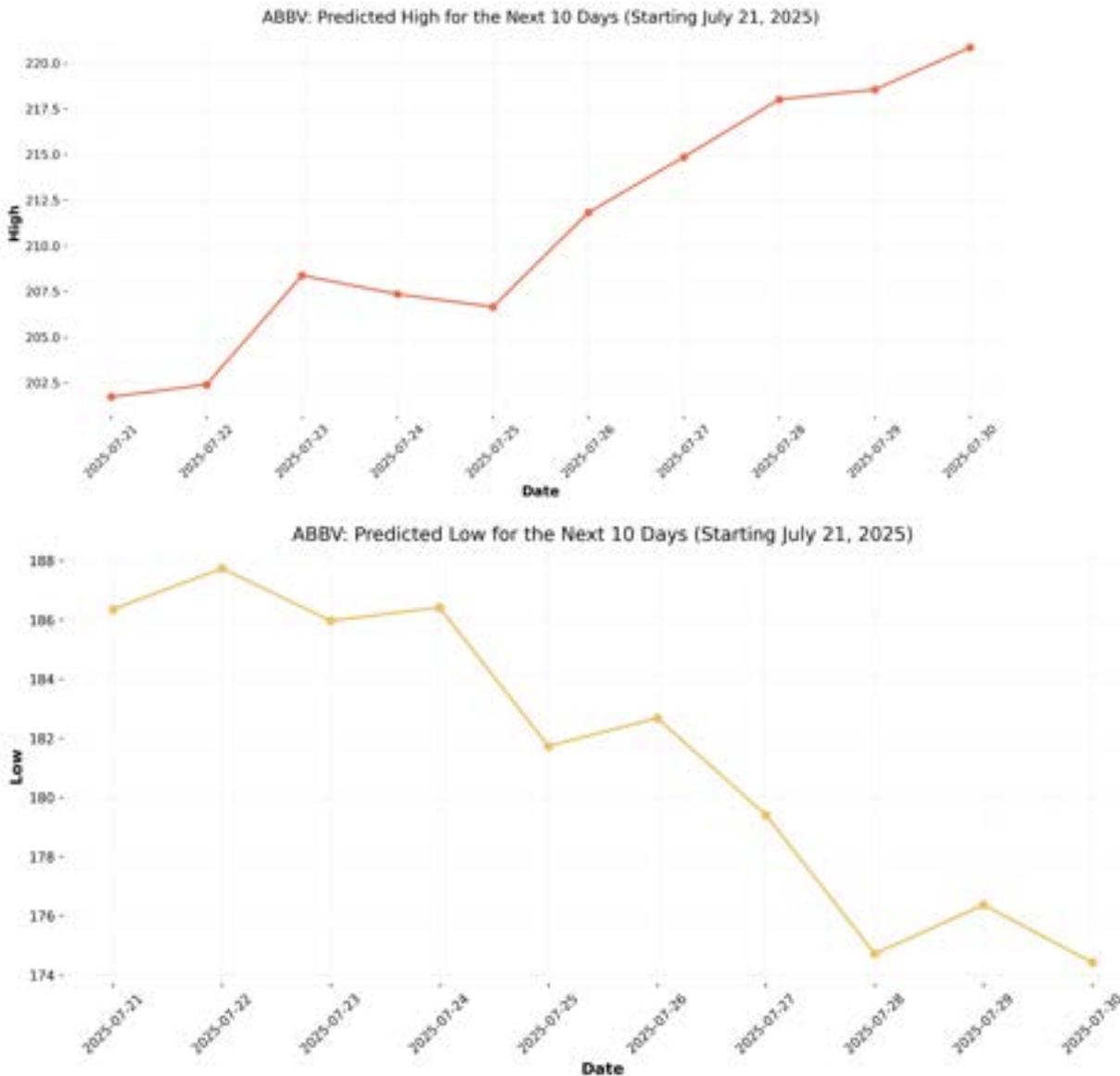
Analyzing the accuracy of the 10 day prediction, model performance varied across companies, with lower prediction error observed for certain stocks during the evaluation window. However, error magnitude differed substantially across firms, reflecting sensitivity to market conditions and price volatility. The accuracy table shows this as the prices are not quite exact and some rather really off, but the table does have some similar values. For example for Apple, the predicted for the first day was 204.06 dollars and the actual was 212.48 dollars which is 8 dollars off. However, some of the models were pretty accurate such as Abbvie only being off by a few

dollars on some days like on 7/23/2025 being off by less than 2 dollars. Eli Lilly was also following the overall movement of the prices but had different values. Furthermore, the accuracy expectations were not too high due to the flaws in the model and the unexpectedness of the market itself.

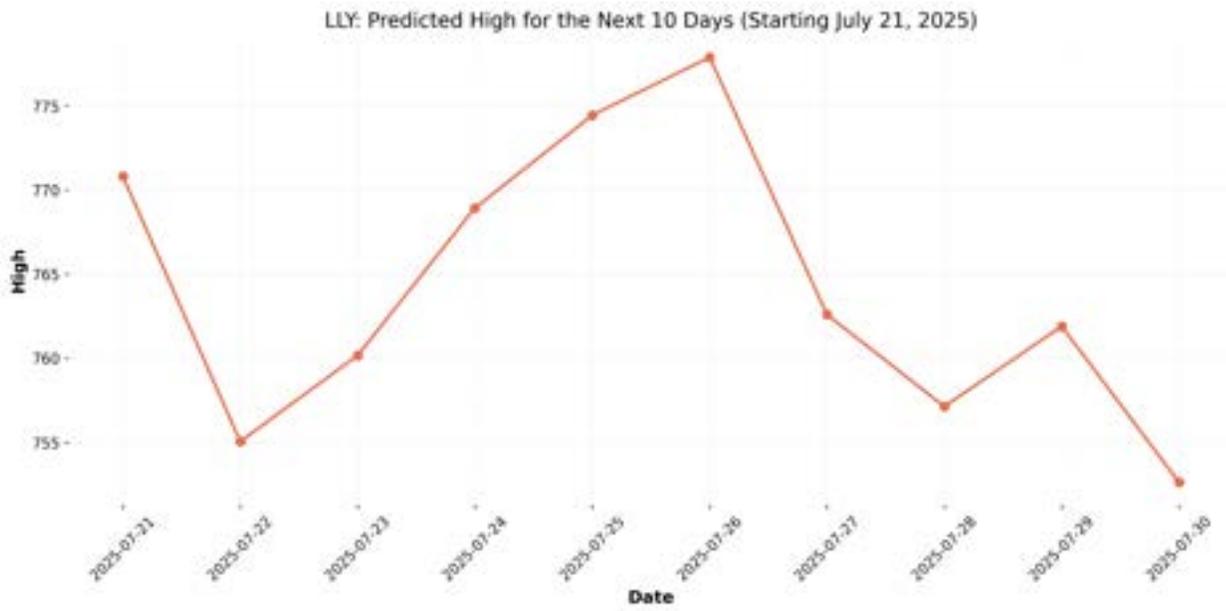
The 100 day prediction model was also included beneath the 10 day model prediction graphs. It shows how the market will move over those 100 days. However, the accuracy of the 10 day models would advise that it will most likely not be that accurate. Also, prediction accuracy decreases when predicting for longer periods of time.

Not only were the closing prices attempted to be predicted, but the highs and lows were also attempted:

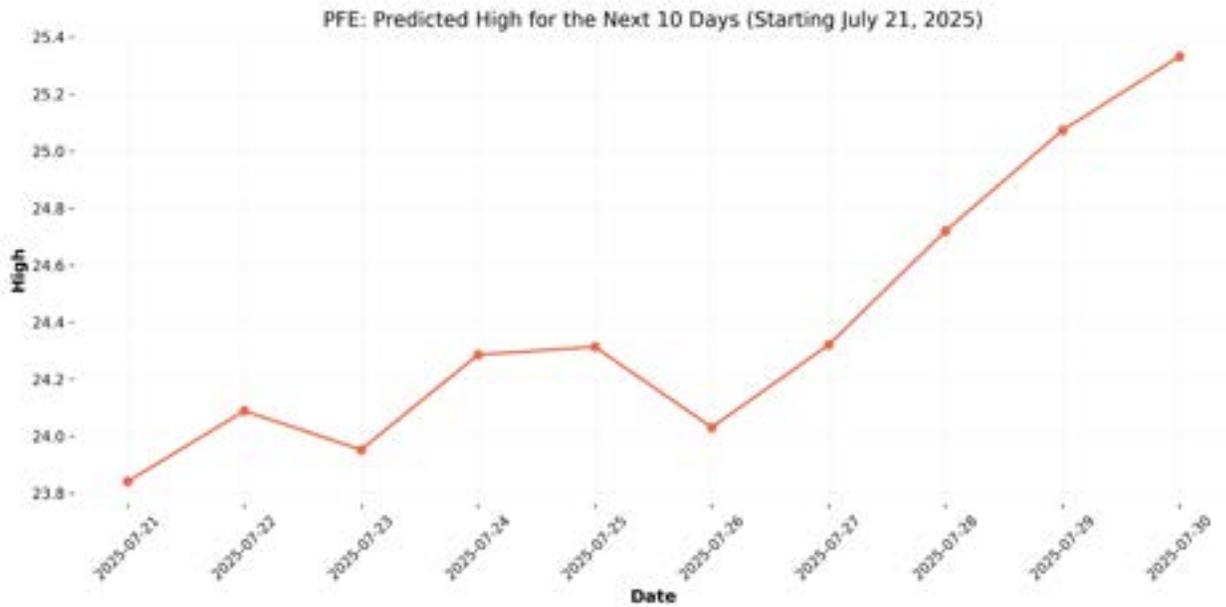
Abbie:



Eli Lilly:

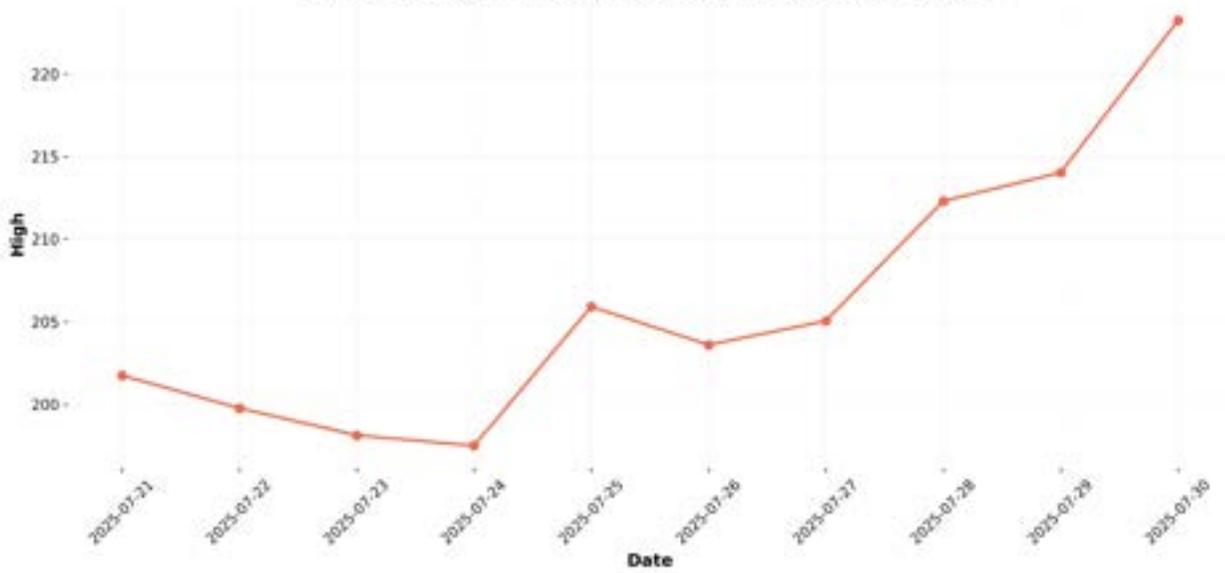


Pfizer:

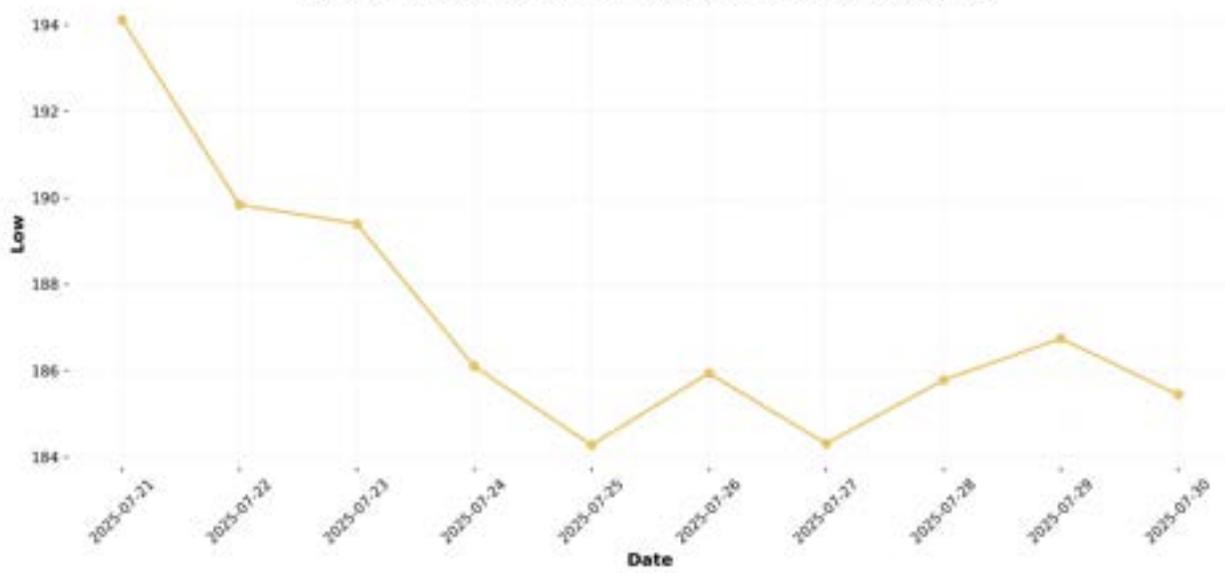


Google:

GOOGL: Predicted High for the Next 10 Days (Starting July 21, 2025)



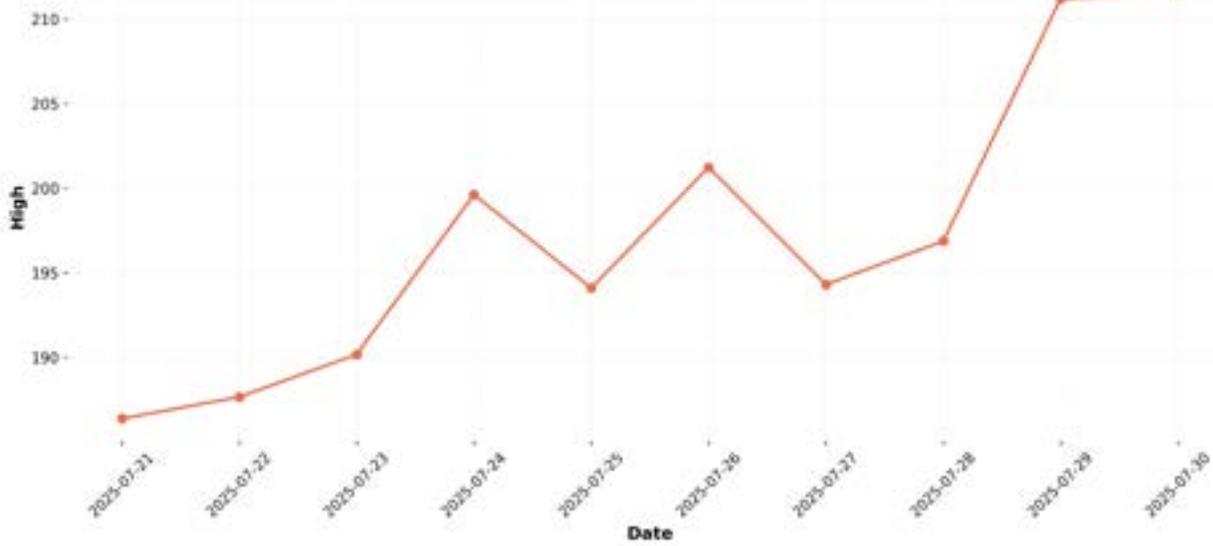
GOOGL: Predicted Low for the Next 10 Days (Starting July 21, 2025)



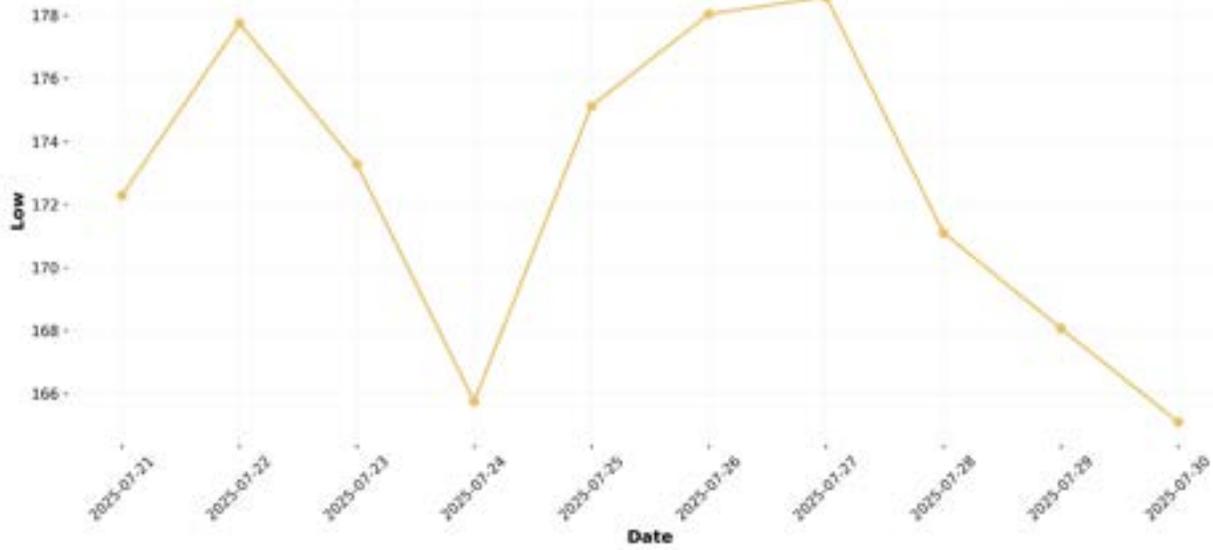
Nvidia:



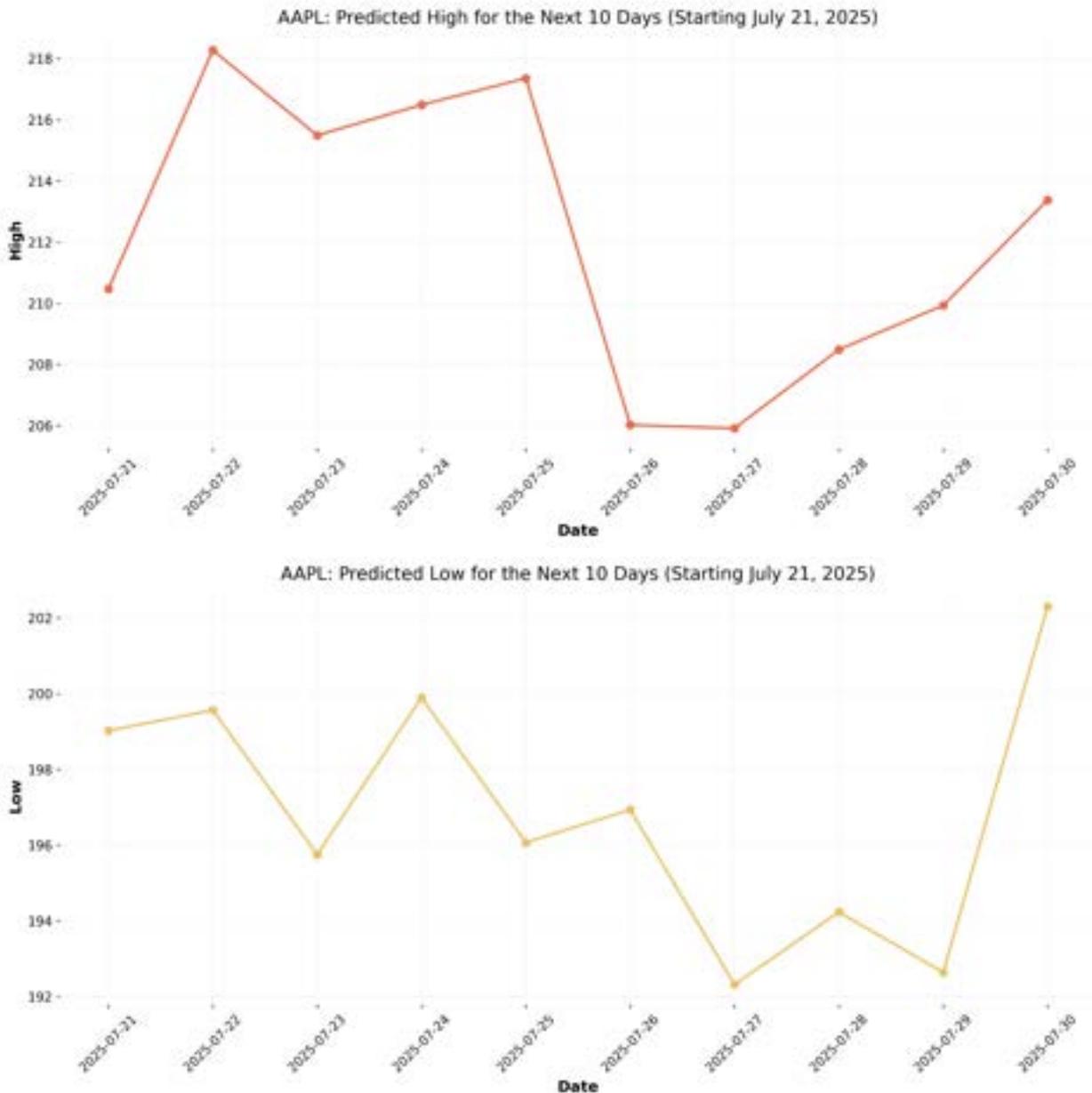
NVDA: Predicted High for the Next 10 Days (Starting July 21, 2025)



NVDA: Predicted Low for the Next 10 Days (Starting July 21, 2025)



Apple:



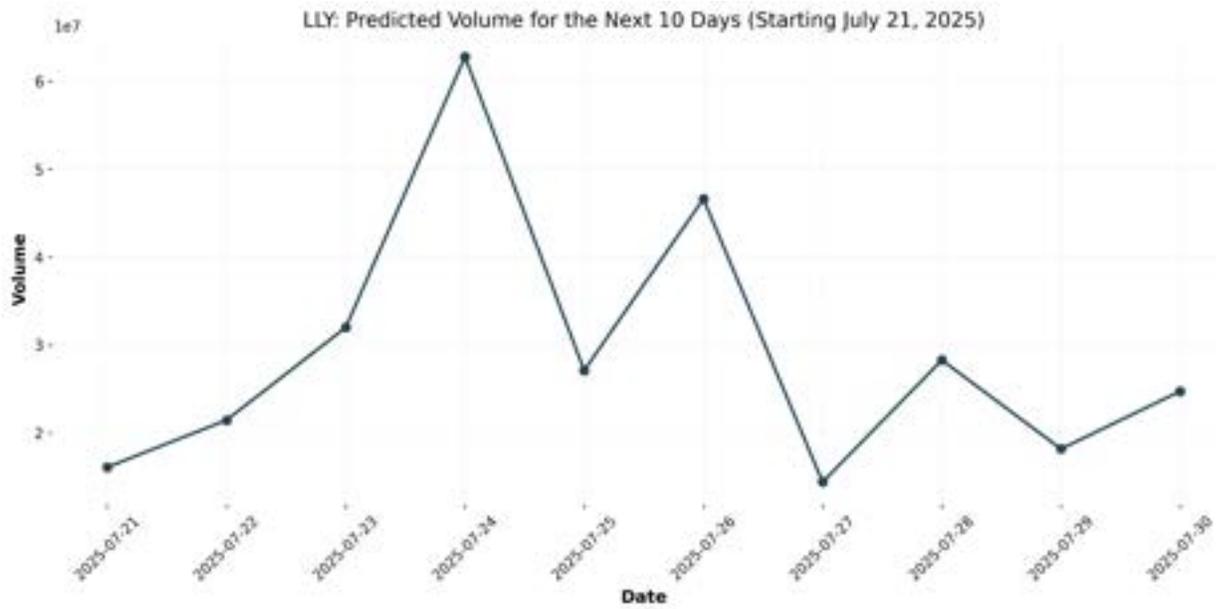
Both the tech industry and the pharmaceutical industry show a wide range in their high and low prices for those 10 days. Some of them are unrealistic in their range and some of the prices were off. For example, Apple's high and low for the 21st of July 2025 was \$215.33 and \$211.19 and the predictions show around \$210 and \$199. However, some were pretty accurate just like in the closing price predictions. Many spikes are also shown in these predictions which is interesting. This is probably due to the factor of past events and patterns which is manipulating the predictions.

The volume was also predicted:

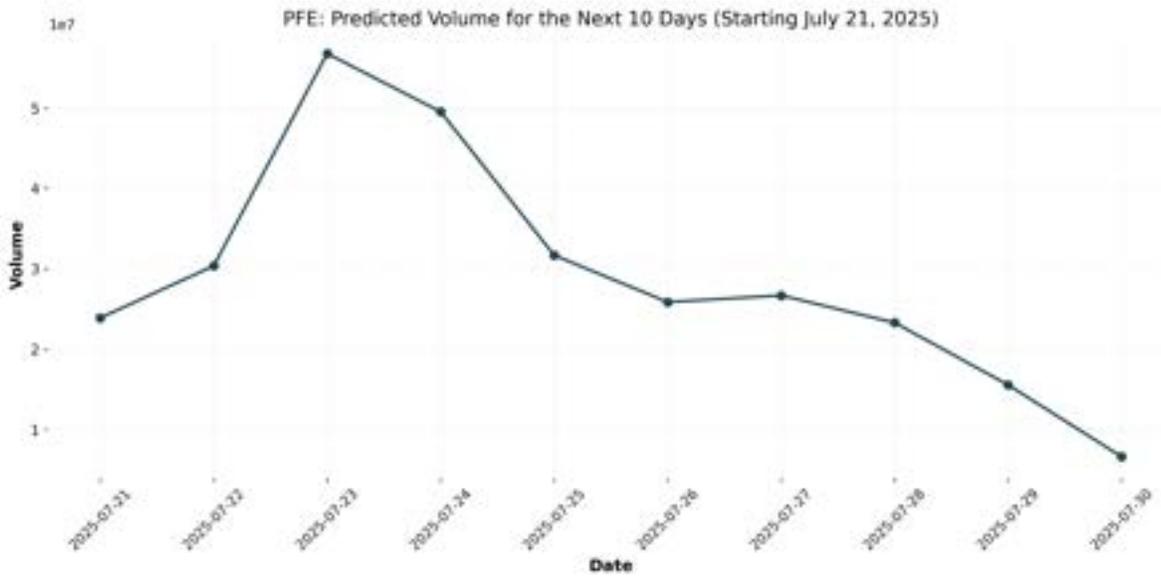
Abbie:



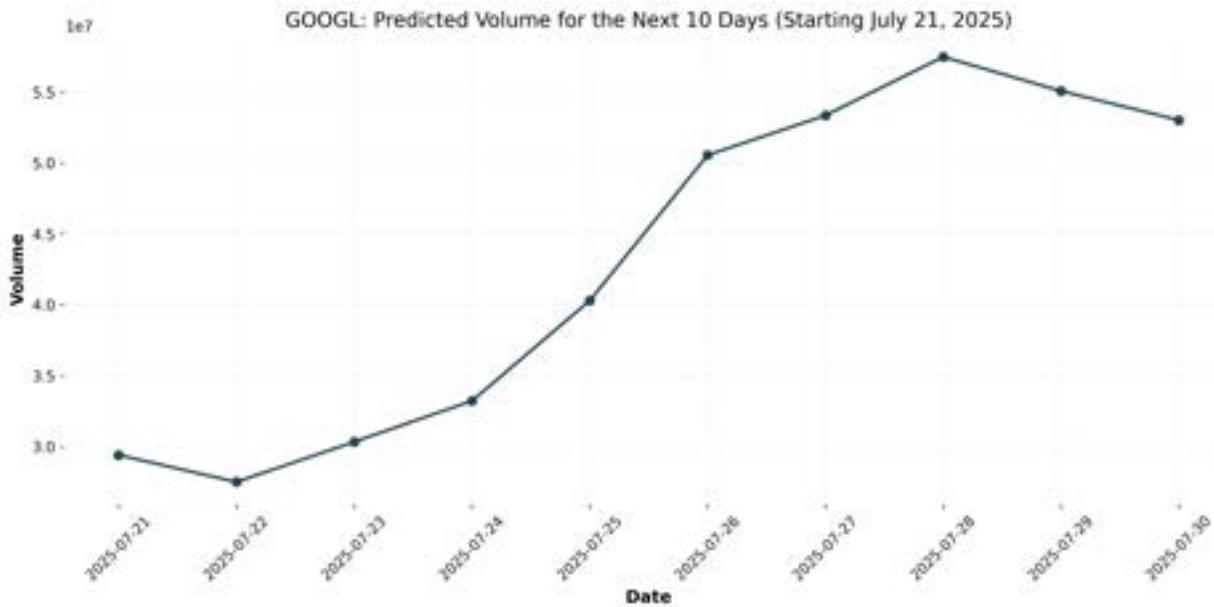
Eli Lilly:



Pfizer:



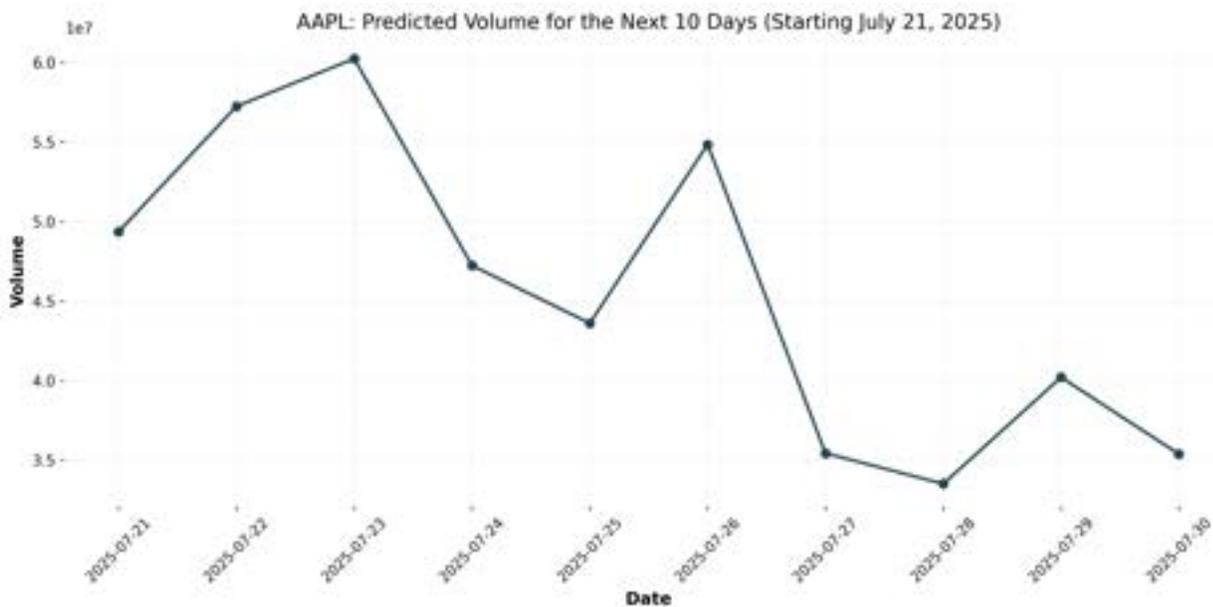
Google:



Nvidia:



Apple:



These are the graphs of the volume predictions of each company. There are a few spikes in some like Eli Lilly, Apple, and Nvidia and some were smoother in their curve such as Pfizer, Abbvie, and Google. In comparing the volumes of each of the industries, there is more volume in the technology industry than the pharmaceutical industry. This is most likely due to their larger investor interest and higher short term trading activity.

The graphs for both sectors are mostly accurate but some of the trends and spikes are a little unrealistic. This data and movement is based on unpredictable factors such as the news which is possible reasons why the model was not accurate here (Beers).

After comparing the predictions from both sectors, it is safe to say that just using historical data and moving averages will not give fully accurate predictions as many other unpredictable factors also go into these movements. When looking at the industries, the technology stocks such as Apple and Nvidia showed sharp rises and falls, which made it harder for the model to predict. However, this does make sense because tech stocks are tied to new innovations, products, and competition. On the other hand, pharmaceutical stocks like Abbvie and Pfizer were more stable at that time even though they had some sudden spikes because of news and drugs. The LSTM models captured the steady patterns in pharma more easily than the tech industry which had more errors.

When looking at individual companies they also showed their own patterns. Apple and Google followed broader market trends and were mostly stable but the predictions were not really accurate probably due to historical trends of those companies. Nvidia was also harder to predict due to the rapid growth of AI. On the pharmaceutical side, Eli Lilly's predictions were fairly accurate and followed the moving average. These results suggest that the model was able to capture short-term trend direction in certain cases, though prediction error remained non-trivial. Abbvie had interesting results; it was following the price movement well by following the rises and falls in prices but the price difference between the actual and the predicted was a little off. Pfizer had more mixed results due to COVID-19 vaccines. This shows that even though companies are within the same sector, they can have different behaviors based on their own stocks and events.

Overall, these predictions show that moving averages are somewhat useful for identifying trends but should not be the only tool used in these predictions. From these results, it is recommended that historical data and moving averages with LSTM models is a good but risky way to get the general pattern in stocks.

Limitations and Next Steps:

Limitations of this research was that the models only used historical price data and moving averages to predict future stock prices (Monfared). They are influenced by many other unpredictable factors such as news, social media, politics, and innovations—especially for the pharmaceutical and technology sectors (Beers). Because these factors were not included, the model's predictions could not be accurate enough to trust it. Even if these factors were implemented, the results would not be fully accurate and trustworthy due to the randomness of the market.

Many people have to predict the market using a ton of real time factors and have succeeded in doing so. Examples of this are Quants, who do this as a job and make a lot of money from the stock market. However, it is also very risky because of the market's randomness, but their chances are higher than most people because they use extremely advanced models, formulas, and real time factors.

Next steps for this research is to add more complex and realistic inputs. Also, adding more advanced models that are able to constantly and consistently run and retrieve real time data such as news would also improve the accuracy of the predictions. However, this project is to

show how historical data and moving averages only can predict or may not predict the future price of a stock. Diving deeper into these kinds of prediction models leads to careers such as quants where a lot of money can be made by accurately predicting the stock market.

Conclusion:

Through this research, it can be concluded that historical trading data and moving averages can help identify general stock price trends, but can't perfectly predict the movement all the time and predict exact prices consistently. LSTM models were used to try predicting two big industries, the technology and pharmaceutical industry. Due to their heavy reliance on innovations and news, without those factors being considered when predicting their stock prices, predictions were inevitably not fully accurate to trust.

This study demonstrates that LSTM models trained on historical price data and moving averages can capture short-term trends in certain market conditions. However, prediction accuracy is limited by the absence of exogenous variables and the inherently stochastic nature of financial markets (Beers).

Acknowledgements:

Manav Patel and Vijay Krishnan (Mentor)

Sources:

- Akana, Tom, et al. *While about Half of Americans Report Owning Stocks Either Personally or Jointly with a Household*.
www.philadelphiafed.org/-/media/FRBP/Assets/Consumer-Finance/Briefs/Why-Some-Americans-Dont-Invest-in-the-Stock-Market.pdf.
- Alberto, Sergio, et al. "The Retail Investor Report." *UMKC School of Law Institutional Repository*, 2023, irlaw.umkc.edu/faculty_works/928/.
- Beers, Brian. "Why Do Stock Prices Change Based on News Reports?" *Investopedia*, 2019, www.investopedia.com/ask/answers/155.asp.
- Mikulic, Matej. "Global Pharmaceutical Industry." *Statista*, 10 Jan. 2024, www.statista.com/topics/1764/global-pharmaceutical-industry/#topicOverview.
- Monfared, Melissa. "Google's Stock Price Prediction ▲ LSTM." *Kaggle.com*, Kaggle, 21 Sept. 2024, www.kaggle.com/code/melissamonfared/google-s-stock-price-prediction-lstm.
- Sayah, Fares. " Stock Market Analysis  + Prediction Using LSTM." *Kaggle.com*, www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm.
- Speights, Keith. "Biotech vs Pharma: What's the Difference?" *The Motley Fool*, 6 Feb. 2026, www.fool.com/investing/stock-market/market-sectors/healthcare/biotech-vs-pharma/.

Velasquez, Francisco. *Tech Volatility Creates the Perfect Environment to “Nibble” on Fintech, Market Veteran Says*. 4 Feb. 2026.
https://finance.yahoo.com/news/tech-volatility-creates-the-perfect-environment-to-nibble-on-fintech-market-veteran-says-172729443.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlmNvbS8&guce_referrer_sig=AQAAADFbk1lcVYkEjHKCxo-WHq534vST_RebSnsN020ilgnO51bqj6iT2GoBKVXLcUbGNhL-AlrCw3ioDQAlyfySjqjELOsyOSA2dHGTSdWuapu-F7kPOqDzJpMVCCcYgn-bNyA3OV-prY0IQeDDEy4j7IVTuSQRK5dANSVTfOhhgqr.