Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Thomas Molinier

20-824-983

Economic Forecasting with Large Language Models: From Stock Pricing to GDP Growth

Master's Thesis

Master's Degree Programme in Applied Mathematics

Swiss Federal Institute of Technology (ETH) Zurich

National University of Singapore (NUS)

Supervision

Prof. Fadoua Balabdaoui, Prof. Huanhuan Zheng

June, 2025

Abstract

Large Language Models are very good at natural language tasks, but their application to quantitative forecasting in economics remains less explored. This thesis explores their abilities on two economic levels: the micro level with stock price forecasting, and macro with GDP forecasting.

For the first part, we integrate an LLM into a formal asset pricing framework, with the goal of predicting institutional investor demand based on stock characteristics, which we then use to calculate an implied price. We then fine-tune it using Group Relative Policy Optimization. While fine-tuning significantly improved the model's demand forecasts, these improvements were not enough to produce price predictions that could consistently beat a simple benchmark.

For the second part, we provide an LLM with a stream of real-time news events and ask it to generate an accurate forecast for annual GDP growth. By benchmarking its predictions against the International Monetary Fund, we find the LLM's forecasts to be very competitive, and often more accurate for several major economies, including the United States, China, and Germany.

Our work shows a contrast. The LLM struggles with the very precise, multi-step reasoning required by the formal asset pricing model. However, it performed better when we asked it to directly synthesize unstructured text into a final quantitative estimate.

Acknowledgements

I would like to sincerely thank Dr. Kelvin J. L. Koa from the National University of Singapore for his availability, continuous support, guidance, and valuable advice throughout this project. I am also grateful to Prof. Huanhuan Zheng for accepting me on the project and offering me this amazing opportunity.

Both Dr. Koa and Prof. Zheng made me feel truly welcome in Singapore, and I learned a great deal during my time in their lab.

I would also like to thank Prof. Fadoua Balabdaoui from ETH Zurich for agreeing to be my official supervisor, and for her kindness.

Contents

I General Introduction **5**

II Stock Pricing using LLM Prediction **8**

1 Introduction **9**

2 Related Works **11**

2.1 Forecasting Methods in Economics 11

2.2 Large Language Models in Finance 11

2.3 Positioning of the Present Study 12

3 Data and Methodology **13**

3.1 Problem statement 13

3.2 Data 14

3.2.1 Stock Characteristics 14

3.2.2 Institutional Holdings 14

3.2.3 Investor Aggregation 14

3.3 Methodology 15

3.3.1 Baseline: Raw LLM 15

3.3.2 Fine-Tuning with Group Relative Policy Optimization . . . 15

3.3.3 From Holdings to Implied Prices 17

4 Empirical Results **20**

4.1 Overall Forecasts Errors 20

4.1.1 Directional Accuracy Analysis (Hit Ratio) 21

4.2 Analysis of Holding-Level Forecasts 23
4.2.1 Evaluating the Pre-trained and Fine-tuned Models 23
4.2.2 Comparative Analysis and Discussion 24

5 Conclusion and Future Work **26**

III GDP Predictions with LLM using News Data **28**

6 Introduction **29**

7 Related Works **31**
7.1 Text as a Macro-Forecasting Signal 31
7.2 From Classical NLP to LLMs 31
7.3 Positioning of the Present Study 32

8 Data and Methodology **33**
8.1 Problem Statement 33
8.2 Data Acquisition and Processing 34
8.2.1 Benchmark Data: IMF World Economic Outlook 34
8.2.2 Input Data: The Wikipedia Event Corpus 34
8.3 Methodology for Generating and Evaluating Forecasts 37
8.3.1 Selection of Target Economies 37
8.3.2 Experimental Design: Information Sets 39
8.3.3 LLM Implementation and Prompt Engineering 40
8.3.4 Evaluation Framework 41

9 Empirical Results for GDP Forecasting **42**
9.1 Overall Forecast Accuracy 42
9.2 The Value of Information: News Events and Recency 53
9.3 Temporal Compliance: Systematic Audit of Reasoning Traces . . . 54
9.4 Discussion 57
9.4.1 Temporal Knowledge Leakage 57
9.4.2 Country–Obfuscation Experiments 58

10 Conclusion and Future Work **60**

IV General Conclusion **62**

A Appendix for Stock Pricing Project **64**
A.1 Language Model Specification 64
A.2 GRPO Hyperparameters 64
A.3 Evaluation Metrics 66
A.3.1 Holding-Level Metrics 66
A.3.2 Price-Level Metrics 66

B Appendix for GDP Forecasting Project **67**
B.1 Language Model Specification 67
B.2 Forecast Prompt Templates 67
B.2.1 Variant 1: $\mathcal{F}^{\mathrm{Mar_t}}(E_{\mathrm{Sep}_{t-1}\to\mathrm{Mar}_t})$ (Recent Window, With Events) 68
B.2.2 Variant 2: $\mathcal{F}^{\mathrm{Mar_t}}(\emptyset_E)$ (Recent Window, No Events) 68
B.2.3 Variant 3: $\mathcal{F}^{\mathrm{Dec_{t-1}}}(E_{\mathrm{Jun}_{t-1}\to\mathrm{Dec}_{t-1}})$ (Earlier Window, With Events) 69
B.2.4 Variant 4: $\mathcal{F}^{\mathrm{Dec_{t-1}}}(\emptyset_E)$ (Earlier Window, No Events) 69
B.3 Audit Prompt for Temporal Leakage Detection 70
B.4 Detailed Audit Alerts 71
B.4.1 Confirmed Temporal Anomaly (1 Instance) 71
B.4.2 Ambiguous Statement (1 of 2) 72
B.4.3 Ambiguous Statement (2 of 2) 73

Part I

General Introduction

Economic forecasting has traditionally relied on structured numerical data, but the recent rise of Large Language Models presents new possibilities for processing and reasoning with complex information. While their success in natural language tasks is well-established, their potential in quantitative and specialized domains like economic forecasting remains a critical frontier of research. This thesis confronts this challenge directly, undertaking a two-part investigation into the practical application of LLMs to predictive tasks at both the microeconomic and macroeconomic levels.

The first part of this work dives into microeconomic forecasting through the lens of asset pricing. This study addresses two primary research questions. First, we investigate the capability of an LLM to learn the complex mapping from fundamental firm characteristics to the portfolio holdings of different investor groups. Second, we evaluate the accuracy of the stock price forecasts yielded by these predicted holdings. This section, therefore, looks into the model's ability to move beyond text and produce precise, numerically-grounded predictions within a defined economic framework.

Shifting from the firm to the nation, the second part of the thesis explores the LLM's capacity for macroeconomic forecasting. Here, the central research question is whether an LLM, when provided with a cleaned stream of real-time news, can generate annual GDP growth forecasts that are competitive with, or even superior to, those from the International Monetary Fund. This investigation further assesses how the time window of the news affects forecast accuracy and introduces an audit procedure to rigorously test the model's temporal compliance, to ensure it does not exploit knowledge of future events.

Viewed side by side, these two studies give us a clear picture of what large language models can and cannot do when we rely on them to forecast economic trends. By testing the models on different scales of analysis (the individual firm vs. the national economy), with different data modalities (structured financial data vs. unstructured news text), and for different purposes (price formation vs. growth prediction), this thesis offers quite a rich evaluation. This approach allows us to identify specific contexts and data types where LLMs are most effective as forecasting tools.

The remainder of this thesis is structured as follows. This general introduction is Part I. Then Part II addresses the microeconomic challenge of stock holding and price forecasting. It begins by reviewing the relevant literature in asset pricing and

machine learning in Chapter 2 before detailing the data and methodology, including the firm characteristics, institutional holdings, and the Group-Relative Policy Optimization (GRPO) fine-tuning approach in Chapter 3. The empirical results for both holding-level and the resulting price-level forecasts are then presented and discussed in Chapter 4. The conclusion for this part is presented in Chapter 5.

Part III then moves on to the macroeconomic level, with Chapter 7 reviewing literature on text-based forecasting. Chapter 8 details the construction of our news event corpus and the experimental design used to test the LLM against an institutional benchmark. This is followed by Chapter 9, which presents a empirical analysis, including the main forecast results, the systematic audit of the model's reasoning, and a discussion of the findings. The conclusion for this part is presented in Chapter 10.

Finally, Part IV synthesizes the findings from both investigations to offer broader conclusions on the role and limitations of LLMs in economic forecasting. The appendices A and B provide supplementary details on the models, hyperparameters, and prompts used in the analyses.

Chapter 1
Introduction

Modern asset pricing research faces a fundamental trade-off between interpretability and predictive power. On one hand, theory-driven models, such as the demand-system framework of Koijen and Yogo [2019], offer a transparent economic mechanism linking stock characteristics to investor behavior, but are often constrained by their own restrictive structures. On the other hand, traditional machine learning models, like gradient boosting or deep neural networks, frequently achieve superior forecasting performance [e.g. Gu et al., 2020, Chen et al., 2024] but operate as "black boxes," making it difficult to understand their decision-making process or trust their predictions in new market regimes. This gap leaves a critical question unanswered: can we build a forecasting model that is both powerful and economically interpretable?

To tackle the problem, we propose and test a hybrid approach meant to bridge the gap between raw text and economic insight. We integrate the flexible, pattern-recognition capabilities of a LLM with the clear economic structure of a demand-based asset pricing system. Instead of using the LLM as a black-box forecaster, we task it with a precise, theory-grounded objective: to predict the quarterly portfolio reallocations of major investor groups based on a standard set of firm fundamentals. These demand predictions are then aggregated and converted into implied stock prices via the market-clearing condition.

Our findings show mixed results. On the first research question, the model shows clear progress: fine-tuning with Group Relative Policy Optimization reduced the Mean Absolute Error of holding-level predictions by as much as 40% for some investor groups compared to the pre-trained baseline. However, this improvement at the demand level does not translate into a clear pricing advantage. The resulting price forecasts fail to consistently outperform a simple naive benchmark, suggesting

that the residual errors in demand estimation can accumulate and overwhelm the pricing signal.

This part of the thesis offers two main contributions. The first is methodological: we propose a modern, theory-grounded approach that uses an LLM to predict investor demand within a market-clearing framework. The second contribution is what we observe from applying this method. We observe that while GRPO fine-tuning is effective at improving holding-level forecasts, these demand-side improvements are currently insufficient to generate accurate price predictions.

The rest of Part I is structured as follows. Chapter 2 positions this study within the literature on demand-system asset pricing and machine learning in finance. Chapter 3 details the data, the LLM fine-tuning process, and the price-derivation methodology. Chapter 4 presents the empirical results for both holding-level and price-level forecasts, and Chapter 5 concludes with a discussion of the implications and directions for future research.

Chapter 2
Related Works

This chapter reviews the key academic literature that provides the foundation and context for our study.

2.1 Forecasting Methods in Economics

Empirical asset-pricing has long linked observable firm characteristics to expected returns and investor behaviour. Early factor models such as Fama and French [1993] treat characteristics as proxies for latent risk factors; more recent "demand-system" approaches explicitly model how heterogeneous investors tilt their portfolios in response to those characteristics. An example is Koijen and Yogo [2019], who show that the small set of size, value, profitability, investment and beta explains much of the cross-section of institutional holdings, providing a transparent mapping from fundamentals to demand. Parallel studies, such as Gu et al. [2020] and Chen et al. [2024], forecast prices directly with machine-learning methods (random forests, gradient boosting and deep nets) often beating linear benchmarks but offering limited economic interpretation. Together, these illustrate two recurring themes: (i) characteristics are powerful empirical signals; (ii) theoretical models improves trust and interpretability.

2.2 Large Language Models in Finance

Large Language Models are quickly moving from generic NLP into mainstream finance, yet mostly for text understanding rather than for theory-based forecasting. Specialised models for finance are now being developed. For example, BloombergGPT,

a model of 50 billion parameters trained on 363 billion finance tokens, is better than general LLMs for tasks like sentiment analysis or answering questions about earnings calls (Wu et al. [2023]). At the same time, open-source projects like FinGPT are making these kinds of tools accessible for applications like robo-advisory and algorithmic trading (Yang et al. [2023]). At the application level, ChatGPT sentiment scores extracted from Wall Street Journal headlines forecast six-month market returns net of transaction costs in Lopez-Lira and Tang [2023]. Yet across these studies the outputs remain sentiment labels or return signals; none embed their forecasts within an explicit asset-pricing framework, leaving the economic mechanisms rather opaque. Tackling that interpretability issue is the goal of the LLM approach proposed here.

2.3 Positioning of the Present Study

The papers above reveal a gap: on one hand, theory-driven economic models are easy to interpret, but they are limited by rigid mathematical forms. On the other hand, LLMs are very flexible at finding patterns in data, but they are often used as 'black boxes' without a clear economic foundation. Our study bridges that gap by using an LLM to predict investor demand from the stock fundamentals, then translating those demand forecasts into implied prices via the market-clearing condition. In contrast to prior LLM studies, the task is fully quantitative and based on a well-defined pricing framework.

Chapter 3
Data and Methodology

Chapter 3 outlines the framework used to address our research questions. We begin by formalising the prediction task and defining the link between characteristics, holdings, and prices. Next, we describe the data set combining stock fundamentals with institutional holdings. Finally, we present the modelling approach, starting with a baseline LLM and extending it with a fine-tuning step, followed by the procedure for translating predicted holdings into implied prices.

3.1 Problem statement

Given five characteristics $x_{k,t}(n)$ for each stock n, we wish to predict the aggregate dollar holding $H_{j,t+1}(n)$ of each broad investor group j one quarter ahead. According to Koijen and Yogo [2019], under standard mean–variance utility, optimal portfolio weights follow an exponential-linear relationship with those characteristics. We therefore ask whether a Large Language Model can learn this mapping directly from data and, through the market-clearing identity, yield an implied price forecast $P_{t+1}(n)$. Success on this task would demonstrate that modern LLMs can absorb structured numerical inputs and produce economically interpretable demand estimates. The entire workflow can therefore be summarised as:

$$\text{characteristics} \longrightarrow \text{predicted holdings} \longrightarrow \text{implied prices.}$$

This set-up leads to two guiding research questions:

RQ 1 – Demand accuracy: Can a Large Language Model learn the mapping from the five characteristics to group-level holdings ?

RQ 2 – Pricing value: Can accurate price forecasts be derived from those predictions ?

Answering these questions requires a matched panel of holdings and characteristics (Section 3.2) and a modelling framework that combines a pre-trained LLM with task-specific fine-tuning (Section 3.3).

3.2 Data

3.2.1 Stock Characteristics

Our core predictors are the following five stock characteristics:

Symbol	Definition (monthly/quarterly frequency)	Raw source
me_t	Log market-equity at quarter-end	CRSP
bm_t	Book-to-market ratio (annual book value aligned to fiscal year-end, carried forward to the next Q1–Q3)	Compustat
$prof_t$	Operating profitability = operating income / book equity	Compustat
inv_t	Asset growth = Δ total assets / lagged total assets	Compustat
$beta_t$	Market beta from a rolling 60-month CAPM regression	CRSP

Variables are winsorised at the 0.5% and 99.5% quantiles and lagged one quarter to avoid look-ahead bias. Market returns come from the value-weighted CRSP index.

3.2.2 Institutional Holdings

Quarterly portfolio disclosures are drawn from Thomson Reuters 13F for U.S. institutions with more than $100 million in listed equity. Each filing is mapped to CRSP PERMNOs via the WRDS-CIK-Link. We retain only common shares (CRSP share codes 10 & 11) and exclude ETFs, ADRs, and closed-end funds. Holdings are measured in dollar value at the report date and reflated to market prices when the filing lag exceeds two business days.

3.2.3 Investor Aggregation

Each filer is assigned to a single group using SIC codes and self-reported type, hence to reduce dimensionality and make fine-tuning computationally feasible, we

collapse the raw filers into seven broad investor groups: (1) Banks, (2) Insurance Companies, (3) Investment Advisors (including hedge funds), (4) Mutual Funds, (5) Pension Funds, (6) Households and non-13F institutions (residuals, defined as the market value not held by the six institutional groups.), and (7) Other 13F institutions. The resulting panel contains, for every (stock, quarter) pair, the dollar holdings of each group and the corresponding stock characteristics.

3.3 Methodology

3.3.1 Baseline: Raw LLM

As a first benchmark we feed each stock's five lagged characteristics directly into a pre-trained Large Language Model without any fine-tuning. The model's hidden layers convert the numeric inputs into a contextual embedding and return a scalar prediction for each investor group's dollar holding. This "out-of-the-box" LLM serves as the naïve reference against which all subsequent enhancements are judged. The forecasting task can be expressed compactly as

$$\widehat{H}_{j,t+1}(n) \;=\; \mathrm{LLM}\Big(\, me_t(n),\, bm_t(n),\, prof_t(n),\, inv_t(n),\, \beta_t(n) \Big), \tag{3.1}$$

where $\widehat{H}_{j,t+1}(n)$ is the dollar holding predicted for investor group j in stock n one quarter ahead. Inputs are the five characteristics defined in Section 3.2.1. The model is prompted with the template 3.1 that states the variable names, their units, and a request for a numeric answer rounded to the nearest dollar.

3.3.2 Fine-Tuning with Group Relative Policy Optimization

To sharpen the out-of-the-box predictions from the baseline model (Eq. 3.1), we apply *Group Relative Policy Optimization* (GRPO) [Shao et al., 2024], a reinforcement learning algorithm that is a memory-efficient variant of Proximal Policy Optimization (PPO) Schulman et al. [2017]. For this task, we fine-tune a *separate*, lightweight adapter for each of the seven investor groups, updating only the final transformer block and the language-model head while freezing the core pre-trained parameters. This fine-tuning uses data from the first quarter of 1980 through the last quarter of

2016, keeping the years 2017 to 2024 for testing. The fine-tuned forecasting rule is:

$$\widehat{H}^{\mathrm{ft}}_{j,t+1}(n) \;=\; \mathrm{LLM}_j\big(me_t(n),\, bm_t(n),\, prof_t(n),\, inv_t(n),\, \beta_t(n)\big), \tag{3.2}$$

where the subscript j indexes the group-specific adapter.

GRPO Mechanism. Unlike standard PPO, which requires training a separate critic (or value) model to estimate the baseline for the reward, GRPO derives this baseline directly from a group of sampled predictions. For each input (i.e., a stock's characteristics), the policy model generates a group of G candidate outputs (holding predictions). A reward is calculated for each, and the advantage is determined relative to the group's average performance. This avoids the computational overhead of a critic model.

Reward Signal. For our specific application, we define a reward signal that converts the absolute prediction error into a smooth, positive score. For each of the G sampled predictions for a stock i, the reward r_i is given by:

$$r_i \;=\; \frac{\omega}{1+\alpha\,|\widehat{H}_i - H_i|}, \tag{3.3}$$

A perfect forecast ($\widehat{H}_i = H_i$) yields the maximum reward ω, while larger errors are penalized hyperbolically.

GRPO Objective. The core of the algorithm is the calculation of the *group-relative advantage*, $\widehat{A}_i$, for each prediction in the group. This is defined as the individual prediction's reward minus the average reward of the entire group:

$$\widehat{A}_i = r_i - \bar{r}, \quad \text{where} \quad \bar{r} = \frac{1}{G}\sum_{k=1}^{G} r_k. \tag{3.4}$$

The policy parameters θ are then updated by maximizing the standard PPO objective function, substituting this group-relative advantage. The loss is:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = \\ -\frac{1}{B}\sum_{i=1}^{B}\Big[\min\big(\rho_i(\theta)\,\widehat{A}_i,\; \mathrm{clip}\big(\rho_i(\theta), 1-\varepsilon, 1+\varepsilon\big)\widehat{A}_i\big) \;-\; \beta\,\mathrm{KL}\big(\pi_\theta(\cdot|s_i)\,\|\,\pi_{\theta_{\mathrm{old}}}(\cdot|s_i)\big)\Big], \tag{3.5}$$

where $\rho_i(\theta) = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{\text{old}}}(a_i|s_i)}$ is the importance weight, $\widehat{A}_i$ is the group-relative advantage from (3.4), and (ε, β) are clipping and KL-penalty hyper-parameters. Full hyper-parameters, such as ω, α, ϵ and β, appear in Appendix A.2.

3.3.3 From Holdings to Implied Prices

A natural extension of our holding-level forecasts is to assess their implications for future stock prices. This can be achieved by employing the market-clearing condition, a fundamental concept in financial economics. The condition states that the total market value of a given stock must equal the sum of all holdings of that stock across all investors. Mathematically, for a stock i at time $t+1$, this is expressed as:

$$P_{i,t+1} \times S_{i,t+1} = \sum_j H_{i,j,t+1} \tag{3.6}$$

where $P_{i,t+1}$ is the stock price, $S_{i,t+1}$ is the total number of shares outstanding, and $H_{i,j,t+1}$ is the dollar holding of the stock by investor category j.

By rearranging this equation, we can derive the stock price:

$$P_{i,t+1} = \frac{\sum_j H_{i,j,t+1}}{S_{i,t+1}} \tag{3.7}$$

This framework allows us to translate our holding forecasts into price forecasts. Using the model's prediction for the holding of each investor category, $\hat{H}_{i,j,t+1}$, we can compute an implied future price, $\hat{P}_{i,t+1}$. The calculation requires a forecast for the shares outstanding at time $t+1$, which is not known at the time of prediction. However, empirical finance literature has established that the number of shares outstanding is a highly persistent time series, especially at a quarterly frequency. Significant changes to share counts are driven by discrete events such as equity issuances or share repurchase programs, rather than continuous fluctuations. Kaplan et al. [2022], for instance, demonstrate the strong predictive power of the lagged share count, showing that it accounts for a substantial portion (81.1%) of the within-firm variation in the next quarter's shares outstanding. They note that forecasting changes is inherently complex, as it involves anticipating these specific, often undisclosed, corporate actions. Given this high persistence and the difficulty of precisely forecasting discrete changes, using the most recently observed value, $S_{i,t}$, as a proxy for $S_{i,t+1}$ is a standard and reasonable approach instudies. Th us,

our price forecast is computed as:

$$\hat{P}_{i,t+1} = \frac{\sum_j \hat{H}_{i,j,t+1}}{S_{i,t}} \tag{3.8}$$

Prompt Template for Holding Prediction

Act as a quant portfolio manager representing {investor_role}. Analyze the following stock fundamentals for stock ID {stock_id} as of {characteristics_date} and output the category's aggregate dollar portfolio holding for that stock in the next quarter.

— **STOCK CHARACTERISTICS EXPLANATION** —
The following five characteristics are key financial metrics used in portfolio analysis:

1. **Market Equity:** Company's market capitalization
- Value: {me} million USD
2. **Book Equity:** Book value of company's equity
- Value: {be} million USD
3. **Profitability:** Operating income before depreciation scaled by book equity
- Value: {profit}
- Measures the company's operational efficiency and profitability
4. **Asset Growth Rate:** Change in total assets (Gat)
- Value: {Gat}
- Indicates how rapidly the company is expanding its asset base
5. **Market Beta:** Systematic risk measure relative to market portfolio
- Value: {beta}
- Shows stock's sensitivity to market movements (beta > 1 = more volatile)

— **CURRENT ANALYSIS TASK** —
Stock ID: {stock_id}
Characteristics Date: {characteristics_date}

INPUT SUMMARY:
Market Equity: {me} million USD
Book Equity: {be} million USD
Profitability: {profit}
Asset Growth Rate: {Gat}
Market Beta: {beta}

Based on these fundamentals, determine the appropriate dollar holding value for this stock in the next quarter. The holding should be a non-negative float representing the value in millions of dollars. Use standard decimal format (e.g., xxxxx.xx) or scientific notation (e.g., x.xxxxe+x) as appropriate.
OUTPUT (ONLY valid JSON): {{"holding_value": <value>}}

Figure 3.1: The structured prompt template provided to the LLM. Placeholders, shown in blue, are replaced with real data for each specific forecasting task.

Chapter 4
Empirical Results

This chapter presents the empirical findings of our investigation, with an analysis structured in two primary stages. We begin in Section 4.1 by assessing the magnitude of the price forecast errors using MAE, MedAE, and MSE. We then refine this price analysis by examining the model's directional accuracy with the Hit Ratio. Then, to understand the source of these price-level results, Section 4.2 evaluates the model's performance on its task of forecasting the absolute dollar holdings for each investor category. Throughout the chapter, we compare the performance of the pre-trained Baseline (Raw) model against the Fine-tuned (GRPO) model. The test period for all results spans from the first quarter of 2017 to the second quarter of 2024.

4.1 Overall Forecasts Errors

In this section, we compare the overall errors of our price forecasts. We use three metrics: the Mean Absolute Error (MAE), the Median Absolute Error (MedAE), and the Mean Squared Error (MSE). These metrics are formally defined in Appendix A.3. We evaluate the performance of the pre-trained Baseline (Raw) model and our Fine-tuned (GRPO) model. We also compare them against a Naive Forecast, which we define as predicting the next quarter's price to be the same as the current quarter's price ($\hat{P}_{t+1} = P_t$).

Table 4.1 shows the results. The first observation is that the Naive Forecast is has the lowest error for all three metrics.

When we compare our two LLM models, we see a clear improvement from the fine-tuning process. The Fine-tuned (GRPO) model performs significantly better

Table 4.1: Comparison of Overall Price Forecast Performance

Model	**MAE**	**MedAE**	**MSE**
Baseline (Raw)	1,238.60	62.37	652,809,589.54
Fine-tuned (GRPO)	588.02	36.80	190,771,540.01
Naive Forecast	19.21	2.63	648,517.91

than the Baseline (Raw) model across all metrics. Specifically:

- The Mean Absolute Error (MAE) is reduced by **52.5%**, from \$1,238.60 to \$588.02.
- The Median Absolute Error (MedAE), which represents the error on a "typical" forecast, is reduced by **41.0%**.
- Most important, the Mean Squared Error (MSE), which heavily punishes large errors, sees a reduction of **70.8%**.

This large reduction in MSE shows that our GRPO fine-tuning not only improves the average forecast but is also much more effective at avoiding the very large, catastrophic prediction errors made by the baseline model. In summary, while predicting price changes remains a very difficult task, the GRPO fine-tuning provides a substantial and consistent improvement over the pre-trained LLM.

4.1.1 Directional Accuracy Analysis (Hit Ratio)

Beyond the overall error metrics, we now investigate the model's ability to predict the *direction* of price changes. For this, we use the Hit Ratio, which measures the percentage of time a model correctly forecasts if a stock price will go up or down. A score above 50% would indicate a useful predictive edge. We perform this analysis on the cohort of persistent stocks seen across the entire test period.

Baseline Performance

First, we look at the performance of the pre-trained Baseline (Raw) LLM. Table 4.2 shows the results for the five best and five worst performing stocks according to this metric. We can see that the hit ratios are very low. The best-performing stock only has a directional accuracy of 25.7%, which is far from the 50% needed to be useful. This result confirms that the raw LLM, in its generalist state, cannot predict the direction of stock price movements.

Table 4.2: Directional Accuracy (Hit Ratio) of the Baseline (Raw) LLM

Best Performing Stocks (Highest Hit Ratio)	
permno	**hit ratio**
13279	25.68%
13844	16.67%
14444	15.32%
46068	15.15%
14536	14.29%
Worst Performing Stocks (Lowest Hit Ratio)	
permno	**hit ratio**
13870	2.97%
42585	3.47%
19502	3.47%
78091	3.85%
49138	3.85%

Table 4.3: Directional Accuracy (Hit Ratio) of the Fine-tuned (GRPO) LLM

Best Performing Stocks (Highest Hit Ratio)	
permno	**hit_ratio_pct**
13279	22.97%
13844	20.59%
46068	15.15%
91083	14.05%
14444	13.71%
Worst Performing Stocks (Lowest Hit Ratio)	
permno	**hit_ratio_pct**
89443	2.49%
91277	3.47%
23026	3.47%
19502	3.47%
83149	3.51%

Fine-Tuned Performance

Next, we evaluate the Fine-tuned (GRPO) model. The results are in Table 4.3. We see that there is no real improvement in directional accuracy. The best hit ratio for the fine-tuned model is 23.0%, which is slightly lower than the best score for the baseline model. The performance for most stocks remains poor.

In conclusion, both the baseline and the fine-tuned models struggle to predict the direction of price changes. Their hit ratios are consistently low and much below the 50% level. This shows that while fine-tuning helps reduce the overall magnitude of the prediction error (as seen in Table 4.1 with lower MAE, MedAE and MSE), it does not give the model the ability to correctly forecast if a stock price will go up or down.

4.2 Analysis of Holding-Level Forecasts

To assess the model's output, we measure how accurately the LLM predicts the dollar holding of a given stock for an investor category in the next quarter. To quantify performance, we use standard regression metrics: the Mean Absolute Error (MAE) and Median Absolute Error (MedAE), which measure the average and median dollar error of the predictions, respectively. Both metrics are defined in the Appendix A.3.1.

4.2.1 Evaluating the Pre-trained and Fine-tuned Models

Our first research question asks whether an LLM can learn the mapping from firm characteristics to the quarterly portfolio holdings of different investor groups. To answer this, we evaluate the accuracy of the model's holding-level predictions, both for the "out-of-the-box" pre-trained model and for the version specialized with GRPO fine-tuning.

Table 4.4 directly compares the performance of the baseline and fine-tuned models for each of the seven investor categories, using MAE and MedAE.

The results clearly demonstrate the impact of fine-tuning. The baseline model struggles with the task, producing substantial MAEs ranging from $6.1 billion for Households to over $9.4 billion for Pension Funds. This confirms that a general-purpose LLM lacks some knowledge required for precise quantitative financial forecasting.

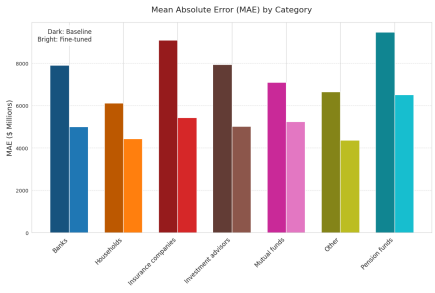
Table 4.4: Comparison of Holding Forecast Performance: Baseline vs. Fine-Tuned (GRPO) LLM

Investor Category	**MAE ($ Millions)**		**MedAE ($ Millions)**	
	Baseline	**Fine-Tuned**	**Baseline**	**Fine-Tuned**
Banks	7905.7	5008.2	254.2	135.4
Households	6122.0	4717.6	236.8	216.9
Insurance Companies	9084.6	5426.4	300.8	192.8
Investment Advisors	7939.7	5010.1	245.5	194.9
Mutual Funds	7100.3	5239.9	368.1	313.1
Other	6655.3	4368.9	259.9	162.5
Pension Funds	9455.9	6514.8	463.8	208.4

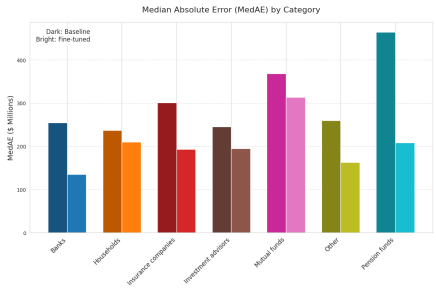
However after fine-tuning, we observe a significant and consistent improvement in predictive accuracy. As shown in Table 4.4 and visualized in Figure 4.1, the MAE and MedAE are reduced for every investor category. The most dramatic improvements are seen for Insurance Companies, where the MAE falls by over 40%, and for Pension Funds, where the MedAE is more than halved. This indicates that the GRPO process successfully trained the models to capture the unique, category-specific patterns in investor behavior.

4.2.2 Comparative Analysis and Discussion

To directly visualize the impact of fine-tuning, Figure 4.1 compares the error metrics of the baseline and fine-tuned models for each category.



(a) Comparison of Mean Absolute Error (MAE) for Baseline vs. Fine-tuned Models.



(b) Comparison of Median Absolute Error (MedAE) for Baseline vs. Fine-tuned Models.

Figure 4.1: Comparison of raw and fine-tuned LLM predictions for both error metrics.

The fine-tuning process consistently reduces the prediction error, as shown in Figure 4.1. The most significant MAE improvements are seen for **Insurance Companies**, **Pension Funds**, and **Investment Advisors**, where the error is reduced by approximately 40%, 31%, and 37%, respectively. For the MedAE, the largest improvement is for Pension Funds. This shows that the GRPO training helped the models learn the specific characteristics of this investor category and reduced the number of very large prediction errors.

In summary, our analysis at the holding level shows that while a pre-trained LLM is not suited for this task on its own, targeted fine-tuning yields meaningful gains in forecast accuracy.

Chapter 5
Conclusion and Future Work

This study explored whether a Large Language Model, guided by economic theory, could forecast institutional investor holdings and, by extension, stock prices.

Our results show that a fine-tuned LLM can learn the relationship between a firm's characteristics and investor demand. The GRPO fine-tuning process significantly reduced holding prediction errors, with Mean Absolute Error falling by up to 40% for some investor groups compared to the raw model. This successfully answers our first research question.

However, this success at the holding level did not translate into valuable price forecasts. When we aggregated the holding predictions to derive market-clearing prices, we faced two major challenges. First, while our fine-tuned model was much better than the baseline, reducing MSE by over 70%, it was still vastly outperformed by a simple naive forecast where the price is assumed not to change. Second, the model showed no ability to predict the direction of price movements, with hit ratios remaining far below the 50% threshold of usefulness.

Looking ahead, several paths could improve these results:

- **Scaling Up the Language Model**
  This study used a relatively small Qwen model. A natural next step is to employ a larger, state-of-the-art foundation model. While larger models possess greater raw capability, fine-tuning them effectively on a specialized numerical task presents a significant challenge, as their vast number of parameters can be resistant to substantial updates from a smaller dataset.

- **Incorporating Higher-Resolution Data**
  Our analysis relied on quarterly holdings and five stock characteristics. Future research could incorporate higher-resolution data, such as daily market

activity, news sentiment scores, or a more granular set of accounting variables. Providing the model with a richer, more dynamic view of a firm's condition could significantly improve its ability to forecast near-term changes in investor demand.

- **Developing Disaggregated, Filer-Level Models**
  To ensure computational feasibility, our approach aggregated thousands of institutions into seven broad categories. A significant next step would be to develop separate forecasting models for each individual 13F filer. This would allow the model to capture the unique strategies and preferences of specific managers, potentially leading to far more accurate demand predictions. However, this approach is currently computationally prohibitive with our method, requiring thousands of separate fine-tuning runs.

Part III

GDP Predictions with LLM using News Data

Chapter 6
Introduction

Gross Domestic Product (GDP) growth is the principal measure of a country's economic health, yet its official figures arrive with significant lags and are subject to frequent revision. This forces policymakers and investors to rely on forecasts to guide decisions in real time. While traditional econometric approaches excel at modeling historical data [e.g. Sims, 1980, Stock and Watson, 2003], they often struggle to incorporate the rich, narrative context of the news flow that comes with major economic turning points. This limitation creates an opportunity for LLMs. Indeed, they are very good at processing unstructured text like news, policy statements, or market reports. Therefore, they could be used to infer how current events might shape future economic activity.

To address this, we develop and evaluate a transparent forecasting pipeline that leverages a LLM to interpret real-time news narratives. By supplying the LLM with a curated, time-stamped stream of news events, we generate annual GDP growth forecasts for nine major economies over the 2007–2024 period. We test four distinct information settings, with different recency and provision of news data. The resulting forecasts are systematically benchmarked against the International Monetary Fund's (IMF) World Economic Outlook (WEO), demonstrating that this approach can produce forecasts that are sometimes more accurate than the institutional benchmark.

This part of the thesis makes three primary contributions:

1. We construct and document a transparent pipeline that translates scraped Wikipedia events into quantitative GDP forecasts, demonstrating a methodology that requires no proprietary data or complex econometric modeling.

2. We provide a systematic benchmark of the LLM's performance against the

WEO across multiple countries and information windows, identifying the conditions under which a news-based model adds the most value.

3. We introduce a procedure for auditing the model's reasoning traces using a secondary LLM, offering a template for systematically verifying temporal compliance in AI-driven forecasting.

The rest of this part is organised as follows. Chapter 7 surveys related work on text-based macroeconomic forecasting. Chapter 8 outlines the problem statement, describes the data pipeline, and details the experimental design. Chapter 9 reports the empirical findings: it provides the overall forecast accuracy, compare different forecasts to determine the value of information, suggest temporal compliance via a systematic audit of the model's reasoning traces, and discusses the broader implications and limitations of the approach. Finally, Chapter10 concludes with directions for future research.

Chapter 7
Related Works

This chapter situates our study at the intersection of text-based macro forecasting and the emerging literature on LLMs.

7.1 Text as a Macro-Forecasting Signal

The idea that news can quantify shifts in macro fundamentals goes back at least two decades. Baker et al. [2016] pioneer an *Economic Policy Uncertainty* (EPU) index built from newspaper archives and show that higher EPU foreshadows lower investment and output growth. Using a broader news corpus, Thorsrud [2020] construct daily topic scores and find that a simple factor of these scores improves quarterly GDP nowcasts relative to benchmark autoregressions. Both papers demonstrate that unstructured text can add predictive content beyond standard numerical indicators, but they rely on bag-of-words or topic models that ignore sentence-level context.

7.2 From Classical NLP to LLMs

The arrival of modern transformer architectures Vaswani et al. [2017] marked a significant technological leap. Unlike earlier methods, models based on this technology can generate contextual embeddings, capturing sentence-level nuance and the relationships between words. This advancement paved the way for the current generation of LLMs, which have been tested directly on forecasting tasks.

The field has since progressed to large-scale generative models, whose capabilities extend beyond text understanding to direct numerical prediction. The sur-

prising effectiveness of these models was highlighted by Gruver et al. [2023], who demonstrated that foundation models like GPT-3 and GPT-4, when prompted with only the numerical history of a time series, could produce forecasts competitive with or even superior to bespoke models explicitly designed for the task.

7.3 Positioning of the Present Study

The literature reviewed above suggests two important points: news-based narratives contain valuable information for macro forecasting, and modern LLMs are capable of translating this raw text into quantitative predictions. Following these ideas, our study presents a detailed case study. We use a clear and reproducible method to test this approach.

Our work specifically focuses on combining three components that are essential for a credible evaluation: (a) reliance on a fully public and reproducible news stream, (b) the strict enforcement of temporal boundaries to prevent knowledge leakage, and (c) a robust benchmark against a strong, real-world institutional forecast. By applying this framework to forecast GDP for nine major economies, we aim to offer a practical and critical assessment of using off-the-shelf LLMs for real-time macroeconomic analysis.

Chapter 8
Data and Methodology

This chapter outlines the data and the experimental framework used to evaluate the effectiveness of Large Language Models in macroeconomic forecasting. We begin by formalizing the prediction task and stating our guiding research questions. We then describe the data pipeline, which transforms raw Wikipedia news events into a structured dataset. Finally, we detail the forecasting methodology, including the experimental design, prompt engineering, and the benchmark models.

8.1 Problem Statement

Traditional macroeconomic forecasting relies on quantitative time-series models (e.g., autoregressions, factor models) that are well-suited to capturing historical patterns in structured data. However, these methods often fail to incorporate the rich, unstructured information contained in the real-time news flow that surrounds major economic turning points. This study tests the hypothesis that a LLM can serve as a bridge, translating news events into quantitative forecasts of macroeconomic indicators.

Specifically, we design and evaluate a pipeline that prompts an LLM with a clean stream of news to forecast the annual percentage change in nominal GDP. The core workflow can be summarized as follows:

$$\text{Raw News Events} \longrightarrow \text{Processed News Events} \longrightarrow \text{LLM Reasoning} \longrightarrow \text{GDP Forecast}$$

This framework leads to three guiding research questions that structure our empirical analysis:

1. **Forecast Accuracy:** Can an LLM, provided with proper news events, produce GDP forecasts that are competitive with institutional benchmarks like the IMF's World Economic Outlook?

2. **Information Value:** How does the accuracy of the LLM's forecasts change with the recency and provision of this event data?

3. **Temporal Compliance:** Can the LLM adhere to strict temporal constraints, and can its reasoning be audited to verify that it is not relying on information beyond its specified knowledge cut-off date?

Answering these questions requires a robust dataset of news events and a carefully designed experimental methodology, which are detailed in the sections that follow.

8.2 Data Acquisition and Processing

Our analysis relies on two primary data sources: (1) official GDP forecasts from the International Monetary Fund, which serve as our performance benchmark, and (2) a novel corpus of macroeconomic news events, which we construct from the public Wikipedia event portal to serve as the main input for our LLM.

8.2.1 Benchmark Data: IMF World Economic Outlook

The target variable for our forecasts is the annual percentage change in nominal GDP. We source the benchmark forecast for each country and year from the April report of the IMF's World Economic Outlook (WEO) database. GDP related variables are often subject to changes. Hence to evaluate forecast accuracy, we decided to use the collect the corresponding final, ground truth values for each country's GDP growth from the 2024 WEO version.

8.2.2 Input Data: The Wikipedia Event Corpus

The core input for our model is a corpus of macroeconomic events constructed from Wikipedia's daily event pages. This process involves a multi-stage data processing pipeline designed to transform raw, unstructured text into a clean, categorized dataset suitable for analysis.

1. Raw Data Scraping We begin by scraping all daily event pages from the English-language Wikipedia for the period spanning January 2002 to December 2024. This initial step yields approximately 90,000 raw text entries. Each entry contains a date, a text description of an event, and, in some cases, a pre-existing categorical tag.

2. LLM-Assisted Preprocessing The raw text is noisy and unstructured. To prepare it for our forecasting task, we employ a two-stage, LLM-driven preprocessing pipeline.

Country Extraction First, to associate each event with one or more economies, we prompt an LLM to identify all countries mentioned either explicitly or implicitly within the event's text. This procedure maps each unstructured text entry to the relevant economies, as illustrated in the prompt template in Figure 8.1.

Prompt Template for Country Extraction

Extract countries mentioned explicitly or implicitly in the following text. Provide a comma-separated list of country names only.

{events_text}

Figure 8.1: The prompt template used to extract countries from each news event.

Category Harmonization Second, we address the issue of inconsistent categorization, as Wikipedia's taxonomy drifts over time. We use a second LLM pass to map the heterogeneous original tags into a consistent set of 16 canonical categories, shown in Table 8.1. For this study, we focus exclusively on events classified as "Business and Economy."

3. Final Event Corpus Once this preprocessing pipeline is complete, we are left with a clean, structured dataset where each "Business and Economy" event is tagged with a date and one or more countries. The properties of this final event corpus are summarized in Figures 8.2, 8.3, and 8.4. This curated data serves as the direct input to the forecasting models detailed in the next section.

Table 8.1: Canonical event categories used for harmonization.

Armed Conflicts and Attacks	Arts and Culture
Business and Economy	Disasters and Accidents
Environment and Health	Games
International Relations	Law and Crime
Media	Miscellaneous News
Politics	Religion
Science	Sports
Transport	Unknown/ContextMissing *or* CategorisationFailed

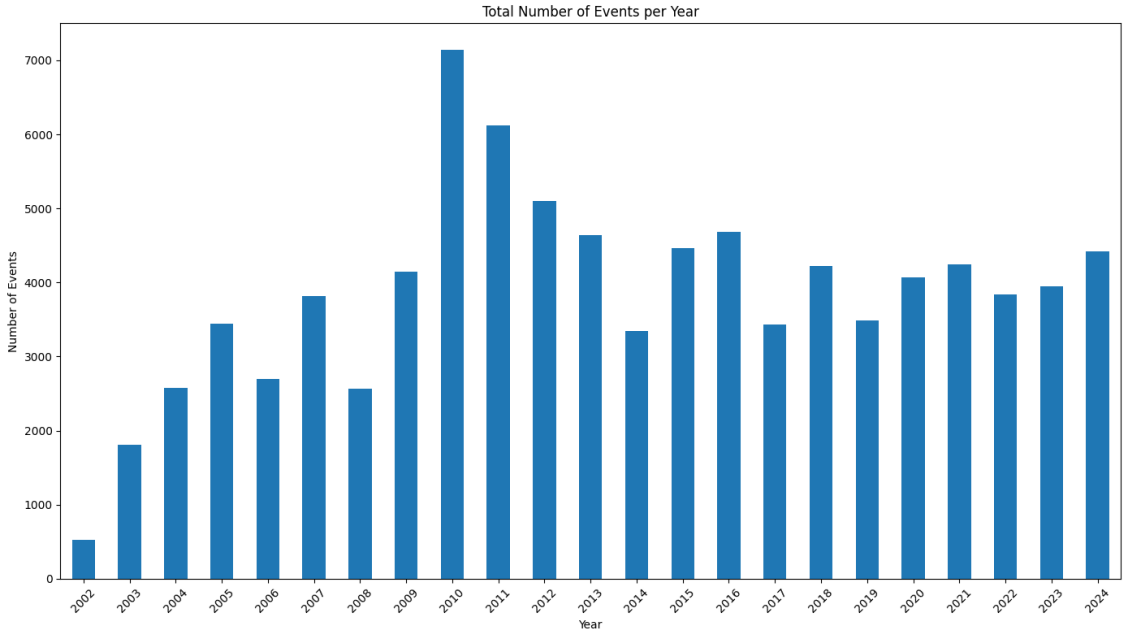


Figure 8.2: Total number of events recorded per year (2002–2024). The surge around 2010–2012 coincides with the heightened Wikipedia editing activity following the global financial crisis and the Arab Spring.

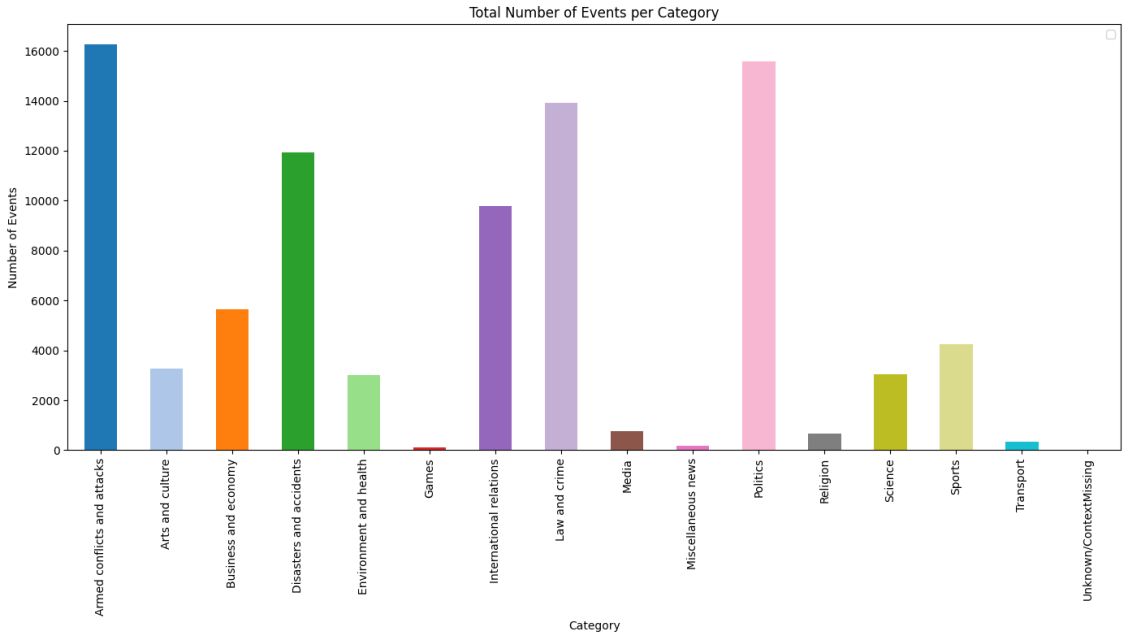


Figure 8.3: Cumulative event counts by final category. Three themes dominate: *Politics*, *Armed Conflicts and Attacks*, and *Law and Crime*. The *Business and Economy* category provides over 5,500 entries.

8.3 Methodology for Generating and Evaluating Forecasts

Having processed the news corpus, we now detail the methodology used to generate and evaluate our GDP forecasts. Our approach involves a controlled experimental design to test the value of news data, a structured prompting technique to elicit forecasts, and a clear set of metrics for performance evaluation.

8.3.1 Selection of Target Economies

Our analysis covers a set of nine major global economies, chosen to provide a diverse and challenging testing ground for our forecasting model. The primary cohort consists of eight of the world's largest and most systemically important economies: the United States, China, Japan, Germany, the United Kingdom, France, Italy, and Canada.

In addition to this group, we include Singapore. The selection of Singapore is motivated by its status as a major open economy and global financial hub, as well

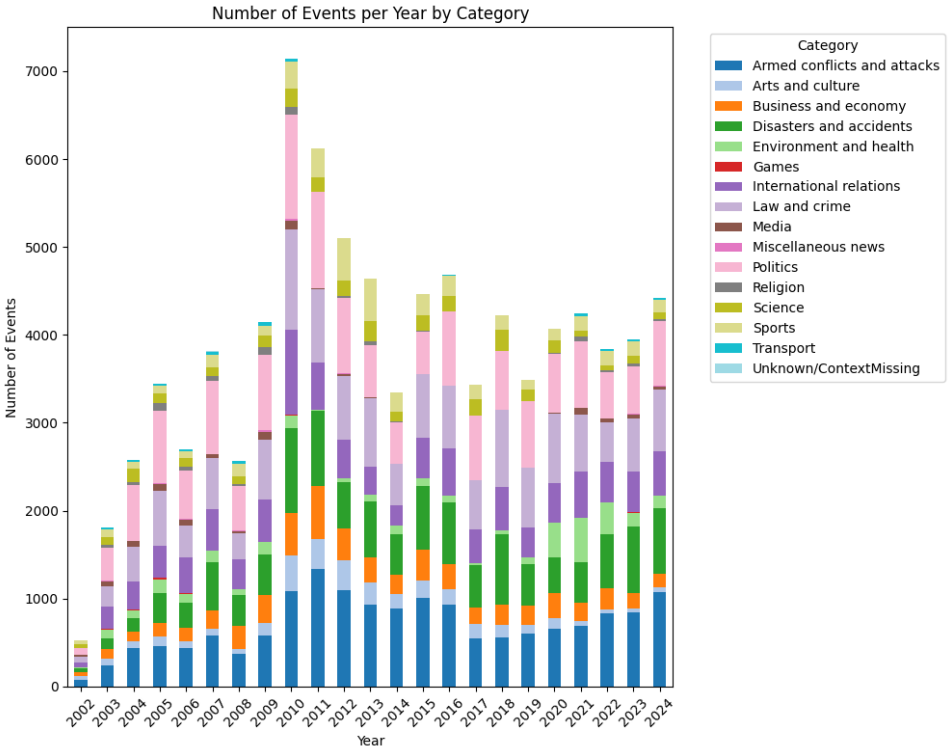


Figure 8.4: Yearly breakdown by category (stacked bars). The plot highlights how the relative salience of categories evolves; for example, *Business and Economy* spikes in 2009–2011.

as its direct relevance to the research environment at the National University of Singapore, where this thesis was developed. It is also interesting to see how our approach generalizes to a more emerging economy such as Singapore.

8.3.2 Experimental Design: Information Sets

To isolate the value of news and its timing, we design an experiment with four distinct information settings for each forecast year t. These settings vary along two dimensions: the LLM's **knowledge cut-off date** and the **provision of news events**. We denote the LLM forecasts with knowledge cut-off date **d** and event set **E** as $\mathcal{F}^{\mathrm{d}}(E)$.

The two knowledge cut-off dates are chosen to provide distinct information sets:

- **End of March of year t ($\mathbf{Mar}_t$):** This aligns closely with the information available to the IMF for its WEO April release. The LLM is strictly instructed not to use any information beyond this date.
- **End of December of year $t-1$ ($\mathbf{Dec}_{t-1}$):** This represents an earlier information set, relying only on knowledge available at the close of the preceding year.

Let us fix a year t. For each cut-off date, we test two scenarios:

- **With Events:** The LLM is provided with a curated list of "Business and Economy" events from our Wikipedia corpus. The event windows are:
  - For the Mar_t cut-off: Events from September of year $t-1$ to March of year t, denoted $E_{\mathrm{Sep}_{t-1}\to\mathrm{Mar}_t}$.
  - For the Dec_{t-1} cut-off: Events from June of year $t-1$ to December of year $t-1$, denoted $E_{\mathrm{Jun}_{t-1}\to\mathrm{Dec}_{t-1}}$.
- **Without Events ($\emptyset_E$):** The LLM is prompted to make a forecast based only on its general internal knowledge up to the cut-off date, without receiving the curated list of events.

This experimental design yields four distinct LLM forecast variants for each country and year, allowing us to systematically test the impact of providing timely, narrative information. Those four variants are $\mathcal{F}^{\mathrm{Dec_{t-1}}}(E_{\mathrm{Jun}_{t-1}\to\mathrm{Dec}_{t-1}})$, $\mathcal{F}^{\mathrm{Dec_{t-1}}}(\emptyset_E)$, $\mathcal{F}^{\mathrm{Mar_t}}(E_{\mathrm{Sep}_{t-1}\to\mathrm{Mar}_t})$ and $\mathcal{F}^{\mathrm{Mar_t}}(\emptyset_E)$.

The motivation for the choice of these two knowledge cut-off date is to see wether including recent events from the beginning of the year we want to forecast yields significant improvements, or if most of the information are already contained in the previous year.

8.3.3 LLM Implementation and Prompt Engineering

The model we use for the forecasting tasks is given in the appendix B.1. For a given country c and forecast year t, we assemble the relevant "Business and Economy" events from our corpus into the structured prompt shown in Figure 8.5. The prompt instructs the model to act as a professional macroeconomist and strictly adhere to the specified knowledge cut-off date.

Prompt: September–March window with events

You are a professional macroeconomist with expertise and data only up to March {t}. Do not incorporate any information or events beyond that date-use only your pre-March {t} knowledge and the events listed below.
You are given a series of economic events related to {c}'s economy that occurred between September {t-1} and March {t}. Based solely on these events and your expertise (with all information limited to before March {t}), provide a forecast of {c}'s GDP percent change for the year {t}. Please include a brief explanation of your reasoning, referencing how these events might influence the GDP outcome.
Important: Ensure that your final output ends with the sentence:
"The forecast for {c}'s GDP change for year {t} is y%," where y is a placeholder for your forecast.
Below is the list of events:
{events_text}

Figure 8.5: The prompt template for a news-augmented GDP forecast. Placeholders are populated with real data for each forecasting task.

The model is instructed to provide its forecast in a specific sentence format, from which we extract the numerical value y using a regular expression. To ensure reproducibility, we use deterministic decoding by setting the model's temperature parameter to 0.

8.3.4 Evaluation Framework

To assess the quality of our LLM-generated forecasts, we evaluate their performance against a strong institutional benchmark using the Root Mean Square Error (RMSE) as our primary metric.

- **Benchmark**: The official forecast for GDP percent change from the IMF's April World Economic Outlook report for year t.
- **Metric**: The **Root Mean Square Error (RMSE)** measures the average magnitude of the forecast errors, with larger errors being penalized more heavily. A lower RMSE indicates a more accurate model. It is calculated across the full 2007–2024 evaluation horizon using the formula:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{8.1}$$

  where $\hat{y}_i$ is the predicted GDP growth for a given year, y_i is the actual ground truth value, and N is the total number of forecasts in the evaluation period.

Chapter 9
Empirical Results for GDP Forecasting

This chapter presents the empirical results of our GDP forecasting experiment. We first analyze the overall accuracy of the four LLM-based models against the IMF benchmark across all nine countries. We then examine the country-level performance in more detail to understand the drivers of these results.

9.1 Overall Forecast Accuracy

Our primary analysis compares the performance of the four LLM forecast variants against the IMF's April WEO projection. The Root Mean Square Error (RMSE) for each model, calculated over the 2007 to 2024 evaluation window, is presented in Table 9.1.

Table 9.1: Root Mean Square Error (RMSE) of GDP Forecasts, 2007–2024. Lower is better. The best-performing LLM variant for each country is in bold.

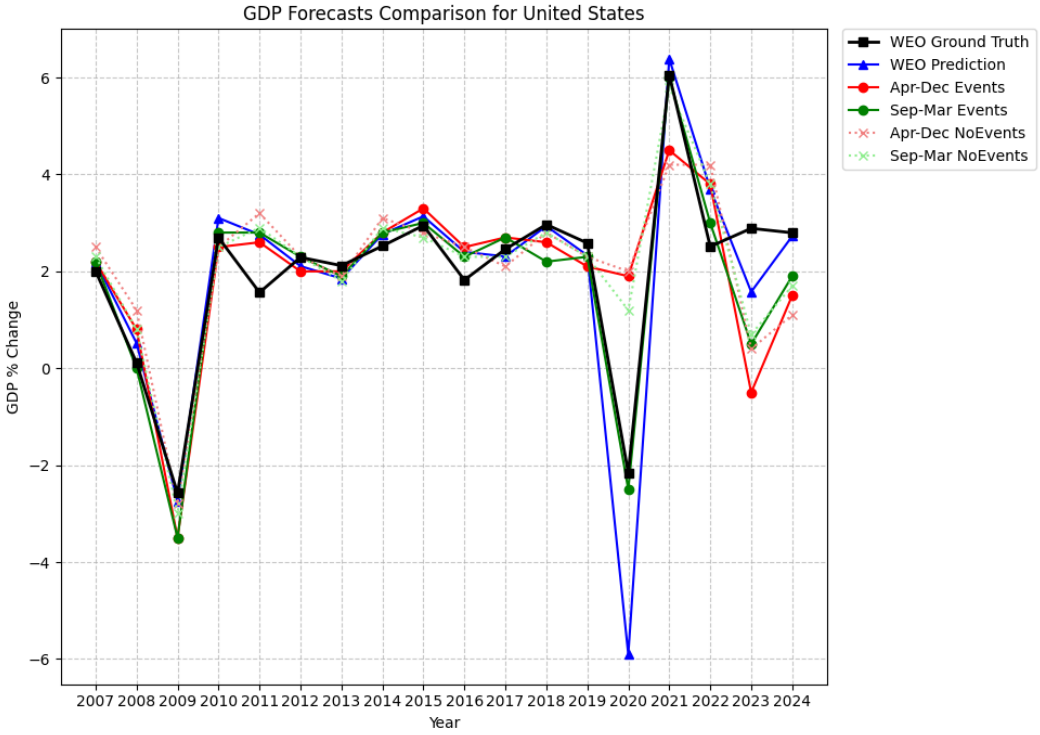
Country	LLM with Events		LLM without Events		IMF WEO (April)
	Jun_{t-1}–Dec_{t-1}	Sep_{t-1}–Mar_t	Dec_{t-1}	Mar_t	
United States	1.44	**0.76**	1.45	1.11	1.05
China	1.45	**1.11**	1.51	1.19	1.29
Germany	1.61	**0.90**	1.77	1.43	1.17
France	2.21	1.46	2.23	2.19	**0.61**
Japan	1.69	**1.10**	1.72	1.43	1.14
Canada	1.77	0.70	1.79	1.59	**0.68**
Italy	2.53	1.90	2.62	2.58	**1.41**
UK	3.02	2.02	2.95	2.95	**1.38**
Singapore	**2.68**	2.95	3.25	2.69	3.70

The results in Table 9.1 reveal several clear patterns. Most notably, the LLM

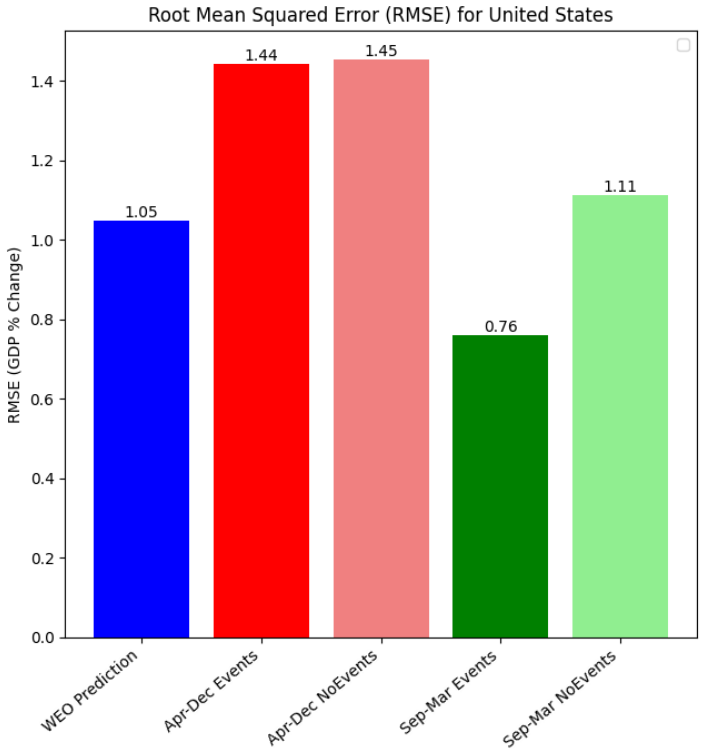
variant provided with the most recent news events ($\mathcal{F}^{\text{Mar}_t}(E_{\text{Sep}_{t-1}\to\text{Mar}_t})$) outperforms the IMF benchmark in four of the nine economies: the United States, China, Germany, and Japan. For Singapore, the LLM with the earlier news window ($\mathcal{F}^{\text{Dec}_{t-1}}(E_{\text{Jun}_{t-1}\to\text{Dec}_{t-1}})$) is the best-performing model overall. Across nearly all countries, providing the LLM with curated news events improves its accuracy, and using more recent news (the Sep–Mar window) consistently yields better results than using older news (the Jun–Dec window).

To better understand these aggregate results, we visualize the full forecast trajectories and the model-by-model RMSE for each of the nine economies. Figures 9.1a, 9.2a, 9.3a, 9.4a, 9.5a, 9.6a, 9.7a, 9.8a and 9.9a plots the forecasted GDP growth against the ground truth, while Figures 9.1b, 9.2b, 9.3b, 9.4b, 9.5b, 9.6b, 9.7b, 9.8b and 9.9b provides a bar chart of the RMSE values from Table 9.1.

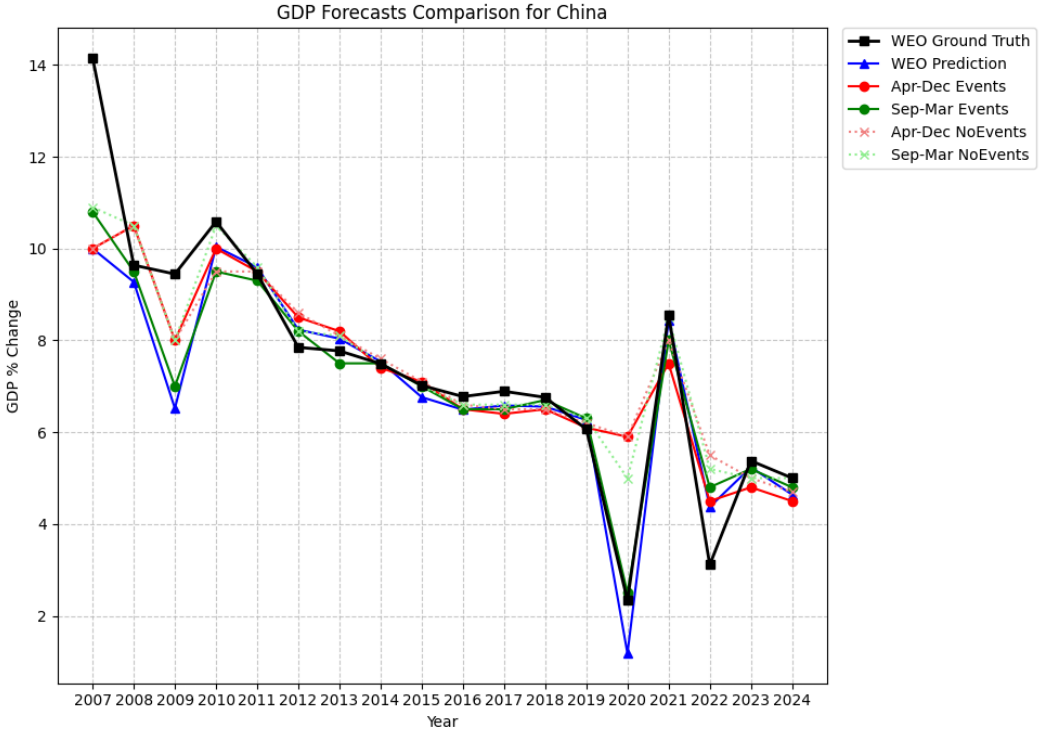
Note on Singapore Data. A key methodological point concerns Singapore, where English-language coverage in the Wikipedia event stream is thinner than for other economies. For the years 2008, 2015, 2016, and 2023, the Sep–Mar window contained no qualifying "Business and Economy" events. In these cases, the $\mathcal{F}^{\text{Mar}_t}(E_{\text{Sep}_{t-1}\to\text{Mar}_t})$ forecast defaults to its "without events" counterpart, $\mathcal{F}^{\text{Mar}_t}(\emptyset_E)$. These substitutions are flagged in the figures.



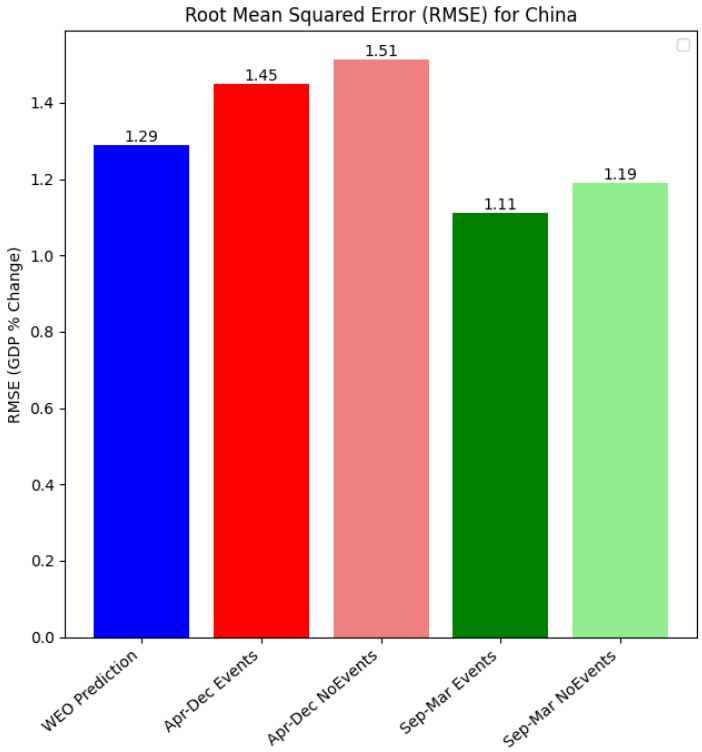
(a) GDP Trajectory vs. Forecasts for the USA



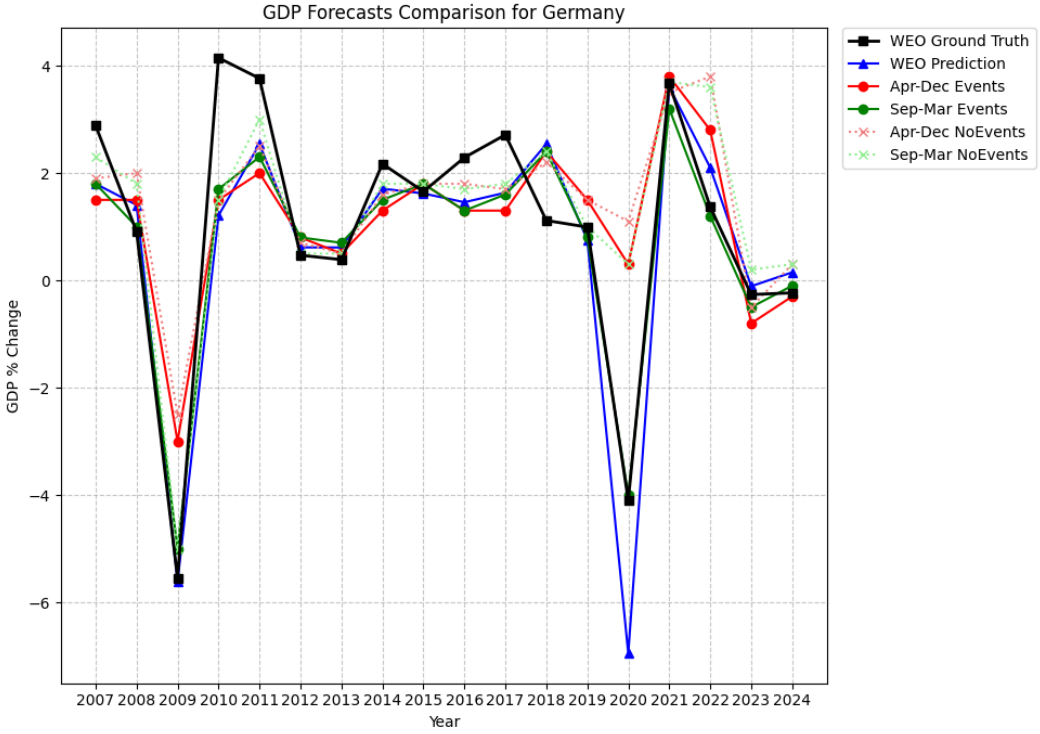
(b) RMSE Comparison of Models for the USA



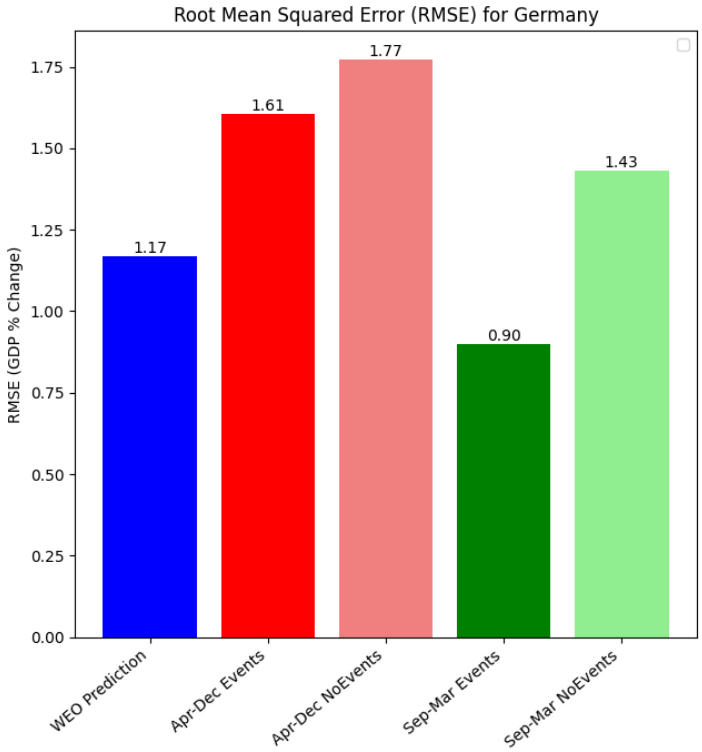
(a) GDP Trajectory vs. Forecasts for China



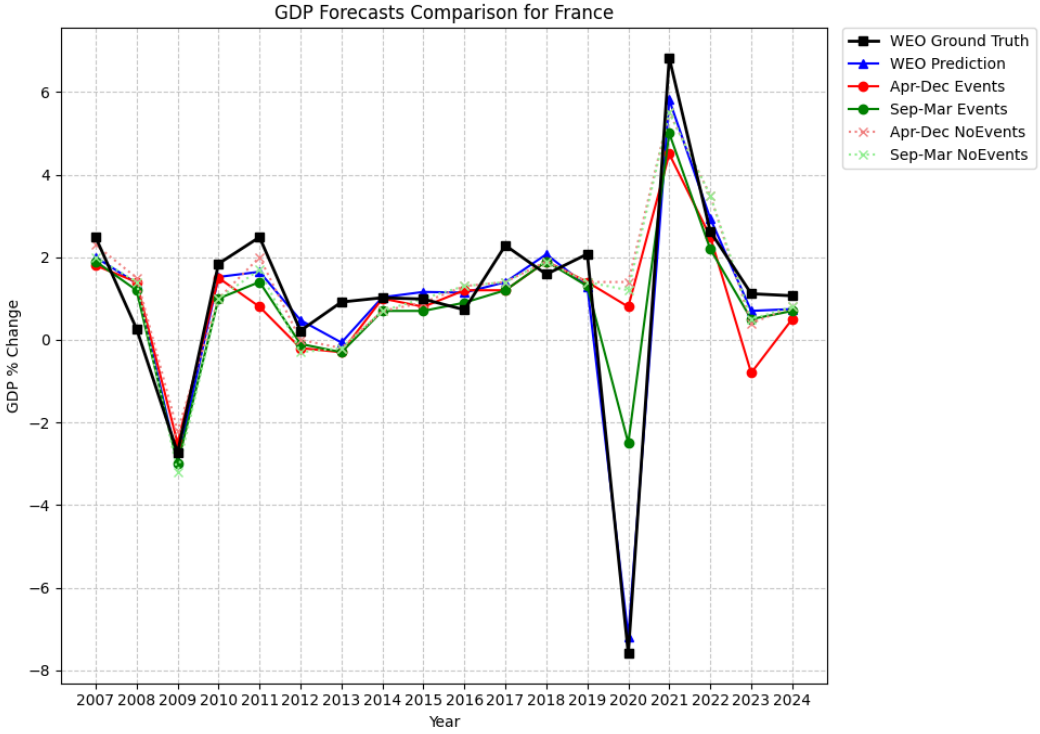
(b) RMSE Comparison of Models for China



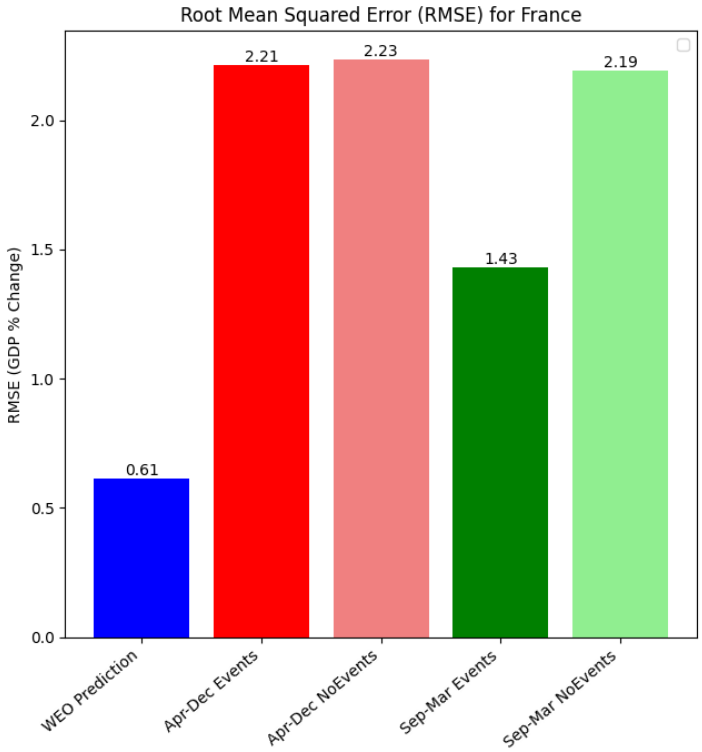
(a) GDP Trajectory vs. Forecasts for Germany



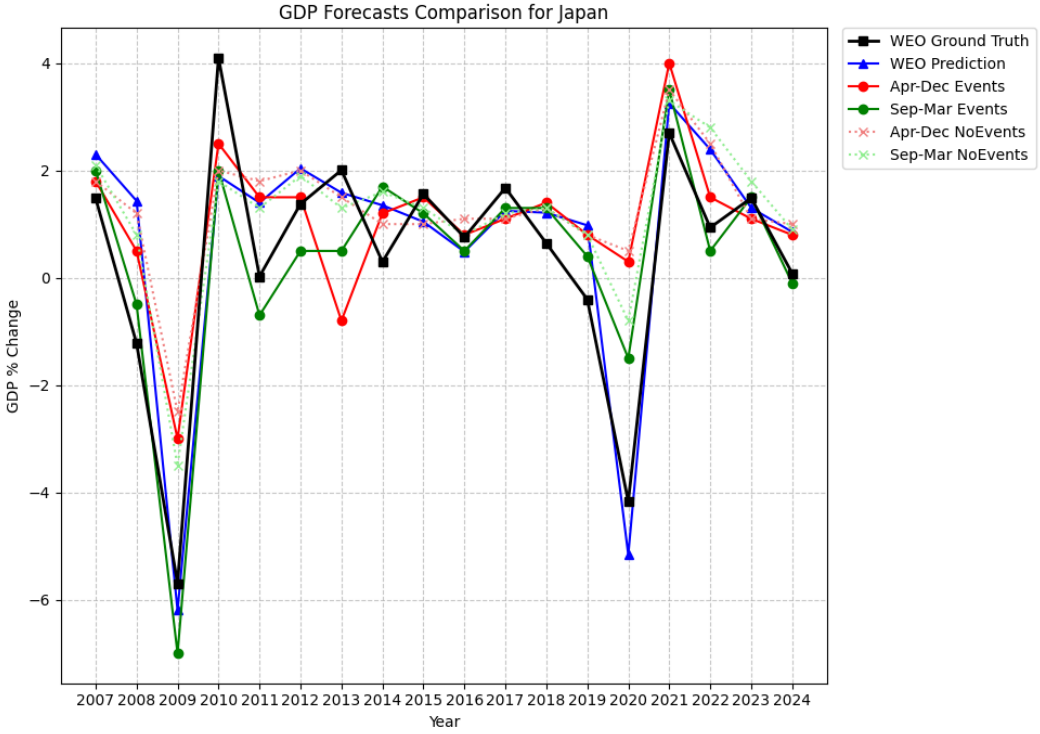
(b) RMSE Comparison of Models for Germany



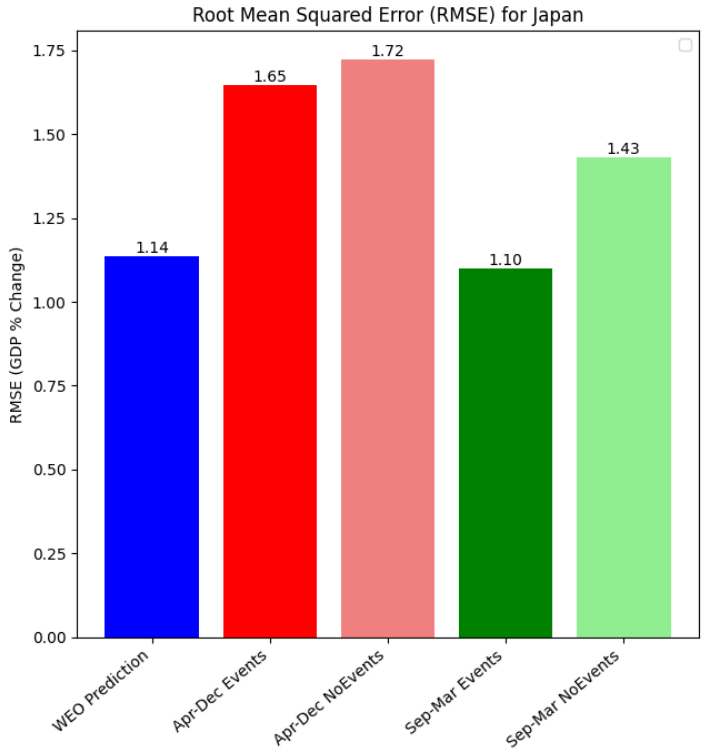
(a) GDP Trajectory vs. Forecasts for France



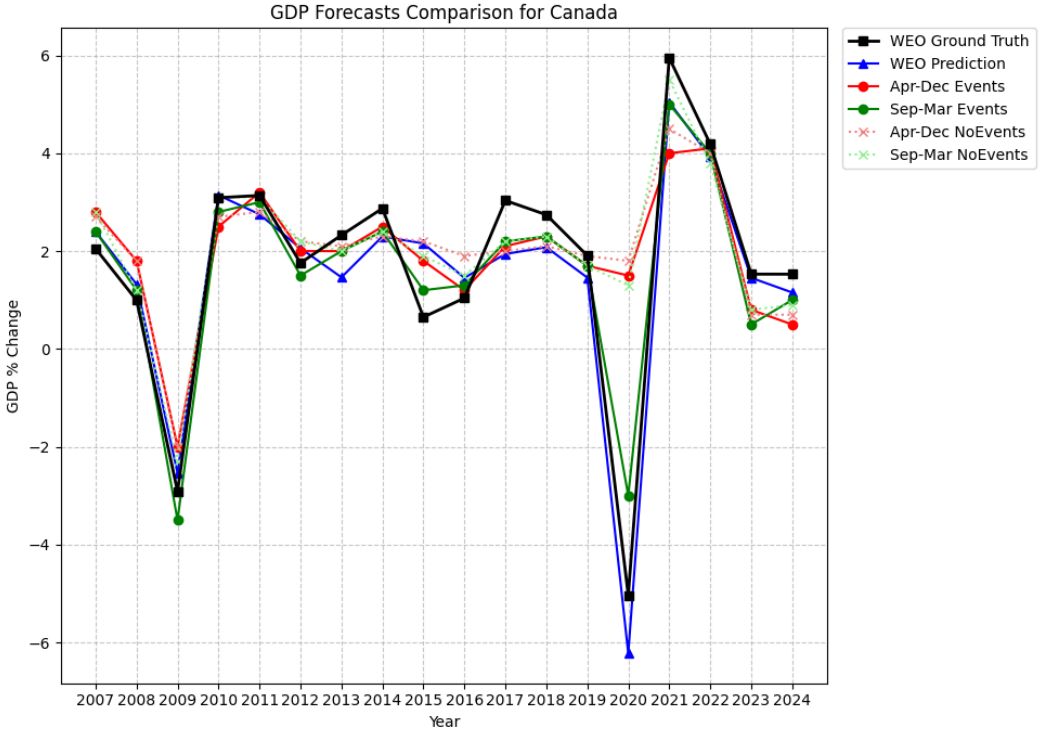
(b) RMSE Comparison of Models for France



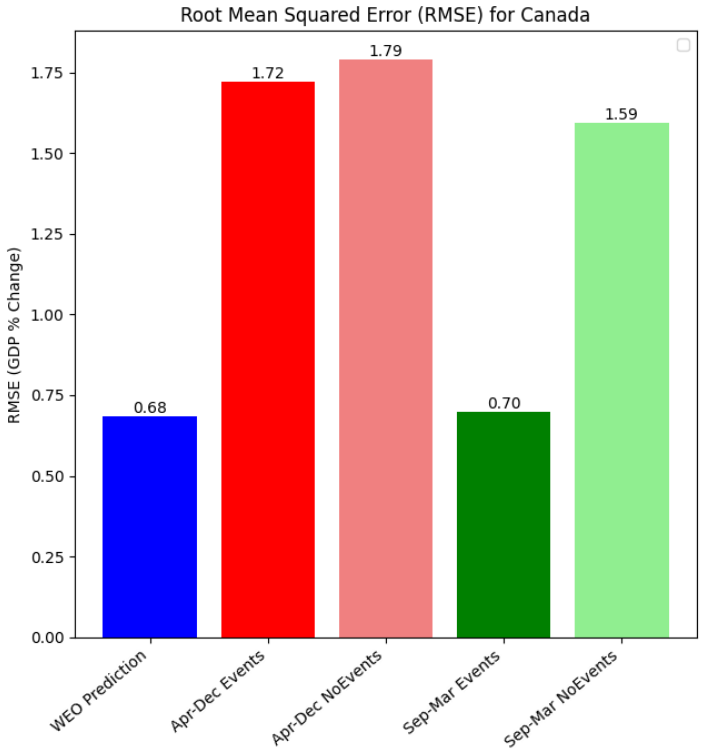
(a) GDP Trajectory vs. Forecasts for Japan



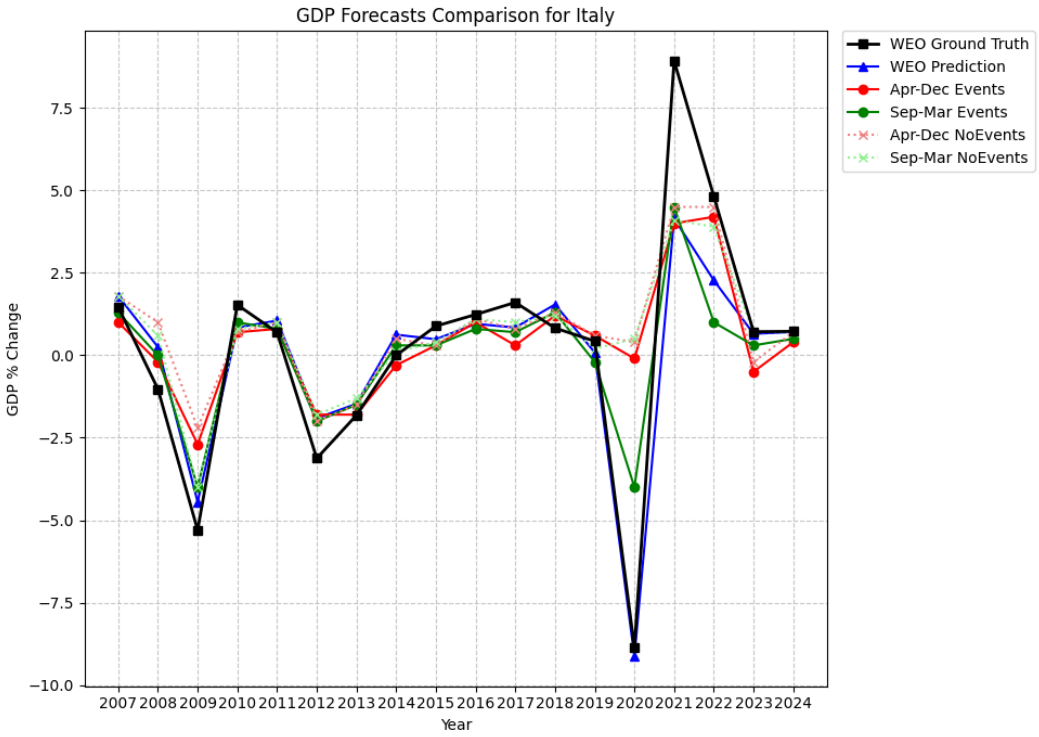
(b) RMSE Comparison of Models for Japan



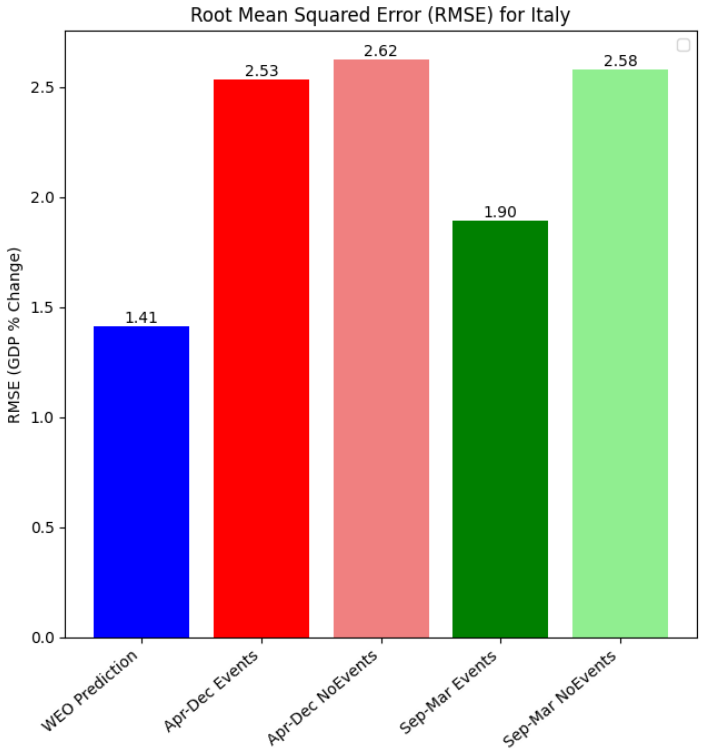
(a) GDP Trajectory vs. Forecasts for Canada



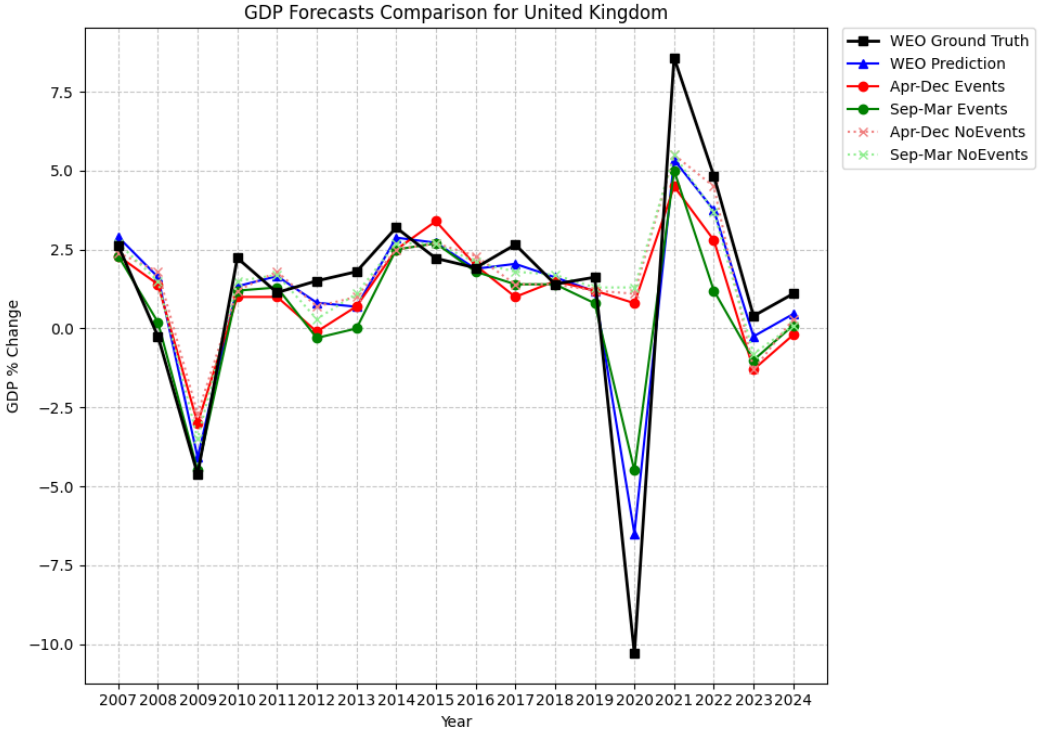
(b) RMSE Comparison of Models for Canada



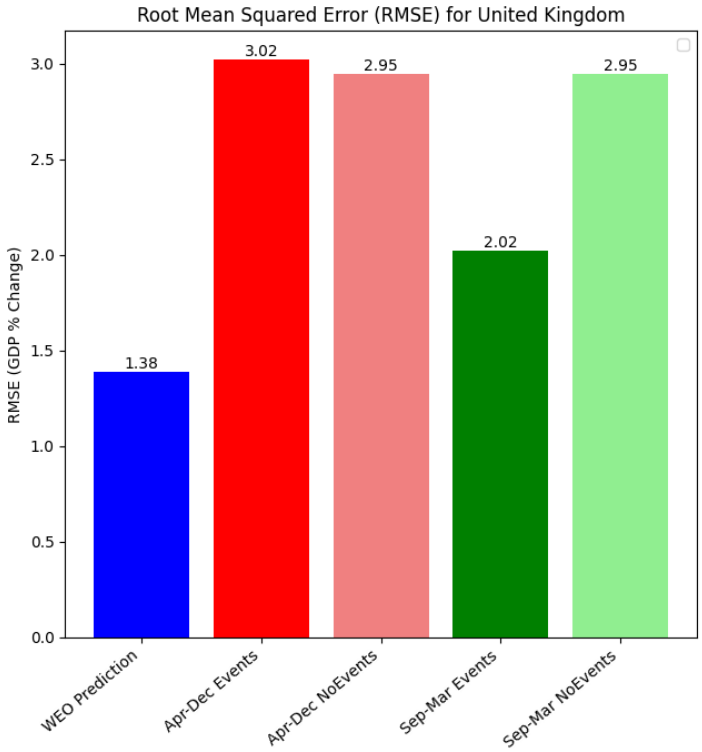
(a) GDP Trajectory vs. Forecasts for Italy



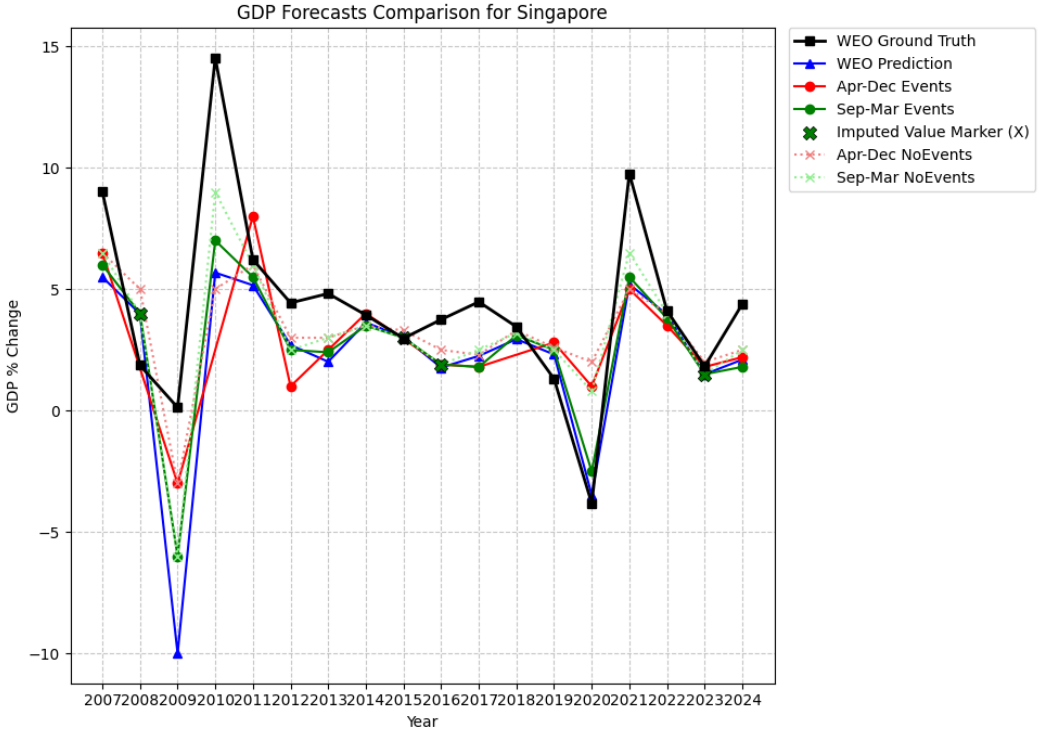
(b) RMSE Comparison of Models for Italy



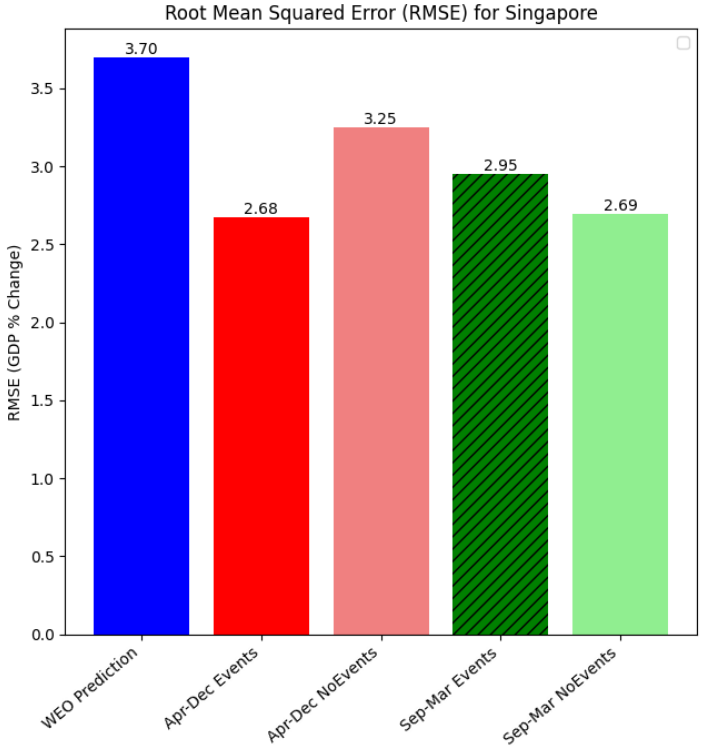
(a) GDP Trajectory vs. Forecasts for the UK



(b) RMSE Comparison of Models for the UK



(a) GDP Trajectory vs. Forecasts for Singapore



(b) RMSE Comparison of Models for Singapore

The visual evidence in the RMSE plots reinforces the findings from the main table and provides additional context. The bar charts clearly illustrate the country-specific nature of the model's performance, showing a distinct advantage for the news-augmented LLM in economies like the US, China, and Germany, while the IMF's forecasts remain more accurate for France, Italy, and the UK.

9.2 The Value of Information: News Events and Recency

The experimental design allows us to isolate the value of providing the LLM with news events and the importance of their recency. Our analysis yields three clear findings on this front:

- **Events matter.** For *seven* of the nine economies (all but the United Kingdom and Singapore), adding the Wikipedia event narrative lowers RMSE relative to the "no-events" prompts for both windows. For the United Kingdom, it improves the results for the closer window, while for Singapore, it improves them for the further one. The effect is clearly visible in Figures 9.1b, 9.2b, 9.3b, 9.4b, 9.5b, 9.6b, 9.7b, 9.8b and 9.9b.

- **Recency helps.** Across all countries but Singapore, the Mar_t window (solid lines) beats the earlier Dec_{t-1} window (dashed lines of the same colour), with **and** without E. The LLM apparently extracts incremental signal from headlines that IMF staff probably also exploit when finalising the WEO April forecast.

- **Recency amplifies the information value of events.** Across the panel the RMSE reduction achieved by adding events is noticeably larger for the recent Mar_t window than for the earlier Dec_{t-1} window. In Figures 9.1b, 9.2b, 9.3b, 9.4b, 9.5b, 9.6b, 9.7b, 9.8b and 9.9b this appears as a wider gap between the dark and light green bars than between the corresponding red bars, indicating that late-winter headlines contribute disproportionately to forecast accuracy once they are surfaced to the LLM.

In sum, this demonstrates that providing a concise, category-filtered news digest allows an off-the-shelf LLM to produce forecasts that are competitive with institutional benchmarks.

9.3 Temporal Compliance: Systematic Audit of Reasoning Traces

While performance metrics provide indirect evidence of temporal compliance, a more direct method involves looking at the reasoning traces produced by the forecasting LLM. We conduct a systematic audit of the textual explanations accompanying all 640 GDP forecasts. This audit employs a secondary LLM specifically prompted to detect any use of post-cut-off information within these reasoning traces (see Appendix B.3 for the audit prompt). The auditor LLM's initial findings were then subjected to human verification for cases it flagged. This review allowed for a nuanced categorization of each instance. The audit process consists in an initial automated scan by the auditor LLM, which categorized traces by printing an alert message together with the associated part of the trace, if the trace was judged to be suspicious. Next, human review re-classified these into more descriptive categories depending on the nature of the observation or the reason for dismissing an alert.

Overall Audit Categorization. Out of 640 audited forecast instances, the initial automated scan by the auditor LLM flagged 45 instances (7.03%) as warranting further review, while 595 (92.97%) were initially assessed as having no anomalies. Following verification of the 45 flagged instances, a final categorization was established, as depicted in Figure 9.10.

Of the 45 instances initially flagged by the auditor LLM, review led to 42 being categorized as "Alert Dismissed" for specific reasons. This left only a very small fraction of cases requiring further consideration: two instances were classified as "Ambiguous Statement" and one instance as a confirmed "Temporal Anomaly". This corresponds to 0.47% of the total forecasts exhibiting an undismissed observation or confirmed anomaly after the full audit process. These three instances can be found in the appendix B.4.

A detailed cross-tabulation of the auditor LLM's initial verdict against the final expert verification outcomes (mapping to the categories in Figure 9.10) is presented in Table 9.2.

Justifications for Dismissed Alerts. For the 42 instances where an anomaly was initially flagged by the auditor LLM but subsequently dismissed by expert review, the justifications were:

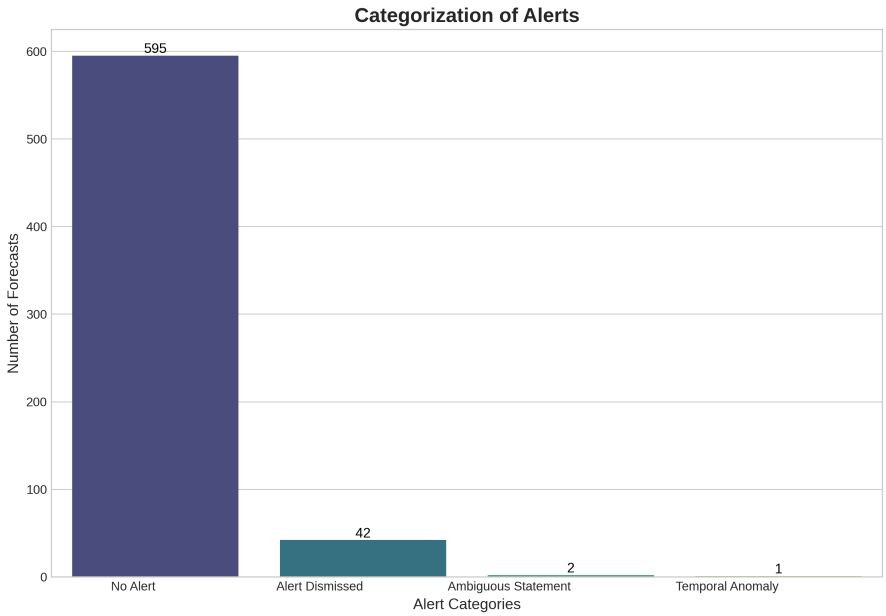


Figure 9.10: Final Categorization of Temporal Audit Alerts after Expert Verification (N=640). The vast majority of forecasts raised "No Alert". Of the 45 initially flagged by the auditor, 42 were subsequently dismissed by expert review.

Table 9.2: Cross-Tabulation of Automated Audit Verdict vs. Final Expert Verification

Auditor LLM Verdict	**Final Audit Status**				**Total**
	No Alert	Alert Dismissed[a]	Ambiguous[b]	Anomaly[c]	
No Anomaly Detected	595	-	-	-	595
Anomaly Flagged	0	42	2	1	45
Total	**595**	**42**	**2**	**1**	**640**

This table shows the verification results for the 640 audited forecasts. Of the 45 cases initially flagged by the automated auditor, human review confirmed only one true temporal anomaly.
[a] *Alert Dismissed* includes cases where the information was verified to be within the permitted window or where the alert was based on a flawed accuracy-as-evidence heuristic.
[b] *Ambiguous* refers to an ambiguity that may be associated with cheating, but is not blatant.
[c] *Anomaly* refers to a confirmed "Temporal Anomaly".

- **Auditor's Accuracy-as-Evidence Heuristic Invalidated:** 23 instances (54.76% of dismissed alerts). In these cases, the auditor LLM appeared to flag forecasts because of their high accuracy. This argument does not itself prove a temporal breach or use of post-cut-off data.
- **Information Verified Within Permitted Window:** 19 instances (45.24% of dismissed alerts). Here, the information questioned by the auditor was confirmed to be explicitly present in the event data provided to the forecasting LLM within its allowed timeframe. For each of those, we found specific elements in the Wikipedia dataset associated with the raised alert.

These justifications suggest that while the automated auditor can be a useful first pass, human oversight is crucial to correctly interpret the nuances of LLM-generated text and avoid misclassifying strong, legitimate inferences as temporal violations.

Influence of Forecast Generation Window on Alert Categories. The distribution of final alert categories was also examined across the different forecast generation windows, as shown in Figure 9.11.

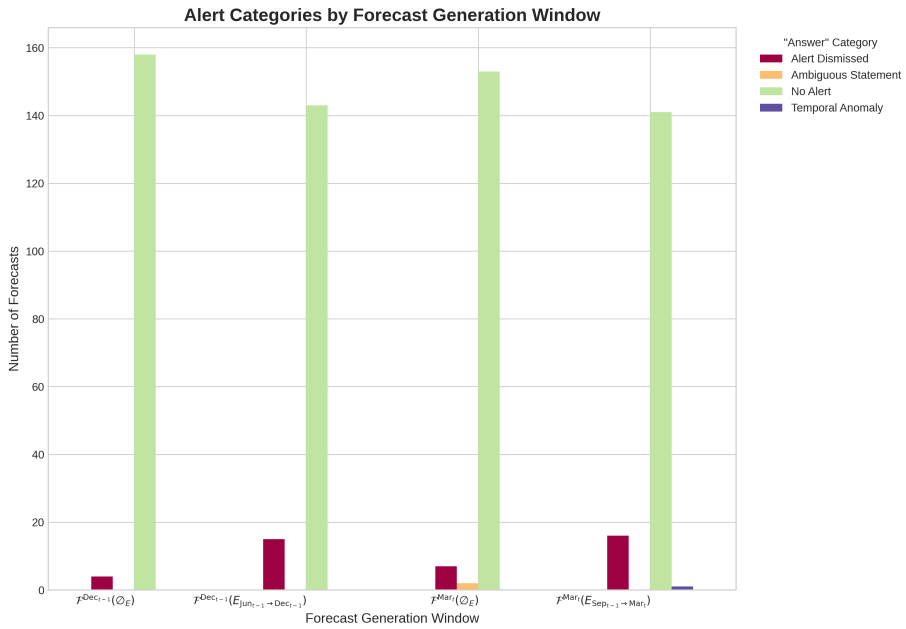


Figure 9.11: Final Audit Alert Categories by Forecast Generation Window. The x-axis denotes the four LLM forecast configurations, varying by knowledge cut-off date and provision of event data. The y-axis shows the number of forecasts in each audit category.

Figure 9.11 indicates that for the earlier Dec_{t-1} knowledge cut-off windows (both $\mathcal{F}^{\text{Dec}_{t-1}}(\emptyset_E)$ and $\mathcal{F}^{\text{Dec}_{t-1}}(E_{\text{Jun}_{t-1}\to\text{Dec}_{t-1}})$), all expert-reviewed alerts were ultimately dismissed. The three instances that remained undismissed (two "Ambiguous Statement" and one "Temporal Anomaly") all originated from the more recent Mar_t knowledge cut-off windows, specifically from the $\mathcal{F}^{\text{Mar}_t}(E_{\text{Sep}_{t-1}\to\text{Mar}_t})$ configuration. While the absolute number of these undismissed cases is extremely small, their occurrence exclusively in the most contemporaneous information setting might suggest that the increased data recency, while beneficial for forecast accuracy (Section 9.2), could also present slightly more complex edge cases for the LLM's temporal grounding.

Overall Audit Interpretation. This systematic audit of reasoning traces indicates a high degree of compliance by the forecasting LLM with the specified temporal constraints. An initial automated scan flagged a small minority of cases (7.03%) for review. Subsequent verification then substantially reduced these concerns, with only 0.47% of all forecasts (3 instances) categorized with an undismissed observation or a confirmed temporal anomaly. The reasons for dismissing initially flagged alerts involved either validating the information source as within the permitted window or correcting for an accuracy-as-evidence argument. While this audit cannot definitively prove the complete absence of any subtle forms of memorisation from the model, it provides no evidence of systematic or widespread exploitation of post-cut-off information in the LLM's generated reasoning. These findings support the cautious optimism that LLMs can be guided to respect temporal boundaries when appropriately prompted and their outputs subjected to careful review.

9.4 Discussion

9.4.1 Temporal Knowledge Leakage

A very important part of our pipeline is the instruction we give to the model "*Do not incorporate any information or events beyond March* t." In practice, the large language models seem to follow our rule most of the time. It is very rare for them to 'hallucinate' and mention data from after the cut-off date. However, we cannot be completely sure that the model has not memorised some later statistics at a deep, technical level in its parameters. Interestingly, forecast accuracy **deterio-**

rates when we move the information window further back ($\mathcal{F}^{\mathrm{Dec_{t-1}}}(E_{\mathrm{Jun}_{t-1}\to\mathrm{Dec}_{t-1}})$ versus $\mathcal{F}^{\mathrm{Mar_t}}(E_{\mathrm{Sep}_{t-1}\to\mathrm{Mar}_t})$, $\mathcal{F}^{\mathrm{Dec_{t-1}}}(\emptyset_E)$ versus $\mathcal{F}^{\mathrm{Mar_t}}(\emptyset_E)$). One benign interpretation is that the model is indeed honouring the temporal boundary, i.e. is yielding less precise forecasts the less contemporaneous information it receives. However, absence of evidence is not evidence of absence: concealed leakage cannot be ruled out without a tightly controlled audit.

Cut-off date check. The base model's public knowledge cut-off is June 2024, so by construction it cannot have memorised the realised 2024 GDP figures that are released after April 2025. Yet the 2024 forecasts display an error profile that is neither unusually good nor unusually bad when compared with earlier years. There are two ways to interpret this similar performance. First, it could be that the model is correctly respecting the cut-off date, so its forecast quality is simply the same as in previous years. The second possibility is that the model's accuracy is just by chance, but it also used a little information from the three future months it was allowed to see (April-June 2024), and this combination accidentally created a result that looks like the past. We only have one year of data, so we cannot be certain. However, the fact that we do not see a large improvement in performance for 2024 at least supports the idea that the model is working correctly.

9.4.2 Country–Obfuscation Experiments

A seemingly simple safeguard against leakage might be to anonymise every country name (e.g., replace France" with Country X"). However, beyond the potential accuracy penalty, two deeper conceptual challenges could also arise:

1. **Shallow masks[1] could prove easy to pierce.** Even after the country name is removed, an LLM could likely still reconstruct the identity from collateral hints such as flagship firms, flagship cities, currency symbols, or well-known geopolitical events. Preventing such an inference would likely require masking an very lagre set of entities and phrases.

2. **Deep masks risk severing the economic graph.** A mask thorough enough to hide all clues could simultaneously erase the web of bilateral trade, supply-chain, and policy links that drive GDP co-movements. Once those

[1]Here a "mask" refers to country obfuscation, such as replacing the actual name by a placeholder like X.

relational signals are gone, the model's ability to reason about growth could be severely diminished.

In short, *light* obfuscation might prove ineffective, while *heavy* obfuscation could be information-destructive. Designing a middle path, e.g., synthetic agent-based worlds with preserved topology, remains an open research problem.

Towards Synthetic World Models. A principled solution would involve constructing a fully self-consistent synthetic universe: agents mapped one-to-one with real countries, maintaining realistic bilateral trade shares, financial flows, and policy reactions. Events would be generated and consumed within this sandbox, preserving network structure while eliminating real-world identifiers. Designing such a simulator is a substantial research undertaking, blending agent-based modelling with controlled language generation.

Chapter 10
Conclusion and Future Work

This study investigated whether a Large Language Model, when supplied with a well-prepared stream of public news events, could produce accurate and trustworthy forecasts of annual GDP growth. By benchmarking four different LLM forecasts against the IMF's World Economic Outlook across nine major economies, we wanted to answer three primary research questions. Our findings provide clear answers to each.

First, regarding **forecast accuracy**, we find that an LLM can indeed produce forecasts that are competitive with, and in several cases superior to, the IMF's institutional benchmark. Specifically, the model variant provided with the most recent news ($\mathcal{F}^{\mathrm{Mar_t}}(E_{\mathrm{Sep}_{t-1}\to\mathrm{Mar}_t})$) achieved a lower RMSE than the IMF for major economies including the United States, China, Germany, and Japan.

Second, concerning the **value of information**, our results demonstrate conclusively that both the provision and the recency of news are critical. Providing the LLM with a curated list of "Business and Economy" events systematically improved its performance over a baseline that relied on its internal knowledge alone. Furthermore, forecasts using news from the first quarter of the target year were consistently more accurate than those using only information from the preceding year.

Third, on the question of **temporal compliance**, our systematic audit of 640 reasoning traces showed a high degree of reliability. With only one confirmed instance of temporal leakage, the audit strongly suggests that LLMs can adhere to strict knowledge cut-off dates when prompted correctly, addressing a key concern in AI-driven forecasting.

In summary, our news-driven LLM pipeline produced GDP forecasts that were often competitive with, and sometimes more accurate than, the IMF's institutional

benchmark. This suggests it can be a useful tool for macroeconomic analysis, particularly when used together with traditional models.

Building on the findings of this thesis, several promising directions for future research could further enhance this framework:

- **Incorporating Additional Event Categories:** Our current model is intentionally constrained, using only events classified as "Business and Economy." A significant next step would be to incorporate other relevant categories, such as "Politics" and "International Relations." Major political events, trade disputes, or geopolitical shocks often have profound economic consequences, and enabling the model to reason over these non-economic narratives could substantially improve its forecasting ability.

- **Building a Synthetic World Model for Rigorous Auditing:** While our audit found minimal evidence of temporal leakage, it cannot entirely rule out subtle forms of knowledge contamination. A more principled solution, as discussed previously, would be to construct a fully synthetic world model. This would involve creating a simulated environment of artificial economies with realistic trade and financial linkages, and then generating a stream of synthetic news events. By testing the LLM in this controlled sandbox, we could be sure that its reasoning is not based on memorized real-world data. This would be a very reliable method for model verification.

Part IV

General Conclusion

This thesis investigated the application of Large Language Models to economic forecasting by tackling two distinct, well-defined problems: one testing the model's role within a formal microeconomic pricing framework, and the other testing its ability to interpret macroeconomic narratives.

In Part I, we tested an LLM within a specific, theory-grounded asset pricing framework. We observed that while fine-tuning with GRPO effectively reduced the prediction errors for investor holdings, the model still struggled to produce accurate price forecasts within our framework.

In contrast, Part II explored the LLM's capabilities in a different domain: macroeconomic forecasting from news text. Here, the research question was whether an LLM could interpret a real-time stream of news events to forecast a country's annual GDP growth. In this task, the model succeeded. By analyzing recent news events, the LLM produced forecasts that were often more accurate than those from major institutional benchmarks like the IMF, demonstrating its strength in turning qualitative, narrative information into a quantitative estimate.

The two parts of this thesis show that an LLM's forecasting ability depends heavily on the task. On one hand, in the stock pricing experiment, where the model was required to generate precise numerical inputs for a rigid formula, its improved demand forecasts were insufficient to yield accurate final price predictions. On the other hand, in the GDP forecasting task, where the model was prompted to directly synthesize unstructured news narratives into a final quantitative estimate, its performance was competitive with a strong institutional benchmark. The effectiveness of the LLM in this thesis was therefore highly dependent on the specific structure of the problem it was asked to solve.

Chapter A
Appendix for Stock Pricing Project

A.1 Language Model Specification

The Large Language Model used for both the baseline and fine-tuning experiments in the stock pricing analysis was a variant of the Qwen2 series. The model was loaded and run locally using the Hugging Face ‘transformers‘ and ‘trl‘ libraries.

- **Model Name:** Qwen/Qwen2.5-3B-Instruct
- **Source:** Hugging Face Model Hub
- **Access Method:** Local inference using ‘transformers.AutoModelForCausalLM‘.
- **Precision:** The model was loaded in full ‘bfloat16‘ or ‘float16‘ precision, without 4-bit or 8-bit quantization, to maintain maximum numerical fidelity for the quantitative prediction task.

A.2 GRPO Hyperparameters

The Group Relative Policy Optimization (GRPO) fine-tuning process was configured with the hyperparameters detailed in Table A.1. These values reflect the final arguments used to launch the training and were held constant across the fine-tuning runs for all seven investor groups.

Table A.1: Hyperparameters for GRPO Fine-Tuning

Parameter	**Value**
Reward Function Parameters (Equation 3.3)	
Reward Scaling Factor (ω)	1000.0
Error Sensitivity (α)	0.001
GRPO Algorithm Parameters (Equation 3.5)	
Clipping Parameter (ε)	0.2
KL-Penalty Coefficient (β)	0.2
Group Size (G)	8
Training Configuration	
Learning Rate	1.0e-6
Training Steps per Category	1000
Batch Size	4
Gradient Accumulation Steps	16
Learning Rate Scheduler	Cosine
Optimizer	Paged AdamW (8-bit)
Random Seed	3407
Generation Parameters	
Maximum Prompt Length (Tokens)	384
Maximum Completion Length (Tokens)	64

A.3 Evaluation Metrics

A.3.1 Holding-Level Metrics

To evaluate the accuracy of the holding-level forecasts, we use two standard regression metrics:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between the predicted holding ($\hat{H}$) and the actual holding (H). It is calculated as $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|H_i - \hat{H}_i|$.
- **Median Absolute Error (MedAE):** The median of the absolute errors. It is less sensitive to large outlier errors than the MAE and provides a measure of the typical prediction error.

A.3.2 Price-Level Metrics

To evaluate the quality of the stock prices derived from the aggregated holding forecasts, we use the following metrics:

- **Mean Absolute Error (MAE)** and **Median Absolute Error (MedAE)** are used similarly to the holding-level analysis to measure the average and median dollar error of the predicted price ($\hat{P}_{i,t+1}$) versus the actual price ($P_{i,t+1}$).
- **Mean Squared Error (MSE)** is the average of the squared errors. By squaring the errors, this metric heavily penalizes large prediction mistakes. A large difference between the MSE and MAE is a strong indicator of the presence of outliers or significant, infrequent prediction failures.

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}(P_k - \hat{P}_k)^2 \tag{A.1}$$

- **Hit Ratio** measures the directional accuracy of the forecasts. It is calculated as the percentage of instances where the model correctly predicted the direction of the price change (i.e., an increase or a decrease) from t to $t+1$. A hit ratio greater than 50% suggests the model has a valuable directional edge, while a ratio near 50% implies performance equivalent to a random guess.

Chapter B
Appendix for GDP Forecasting Project

B.1 Language Model Specification

All GDP forecasting and auditing tasks were performed using a model from the Google Gemini family, accessed via the Google Generative AI Python SDK.

- **Model Name:** gemini-2.0-flash-thinking-exp-01-21
- **Provider:** Google
- **Access Method:** Google Generative AI SDK (‘google.generativeai‘)
- **Configuration:** Temperature set to ‘0.0‘ for deterministic outputs.

B.2 Forecast Prompt Templates

Four distinct prompt templates were used to generate the LLM forecasts, corresponding to the four information sets in the experimental design.

B.2.1 Variant 1: $\mathcal{F}^{\text{Mar}_t}(E_{\text{Sep}_{t-1}\to\text{Mar}_t})$ (Recent Window, With Events)

Prompt: September–March window with events

You are a professional macroeconomist with expertise and data only up to March {t}. Do not incorporate any information or events beyond that date-use only your pre-March {t} knowledge and the events listed below.
You are given a series of economic events related to {c}'s economy that occurred between September {t-1} and March {t}. Based solely on these events and your expertise (with all information limited to before March {t}), provide a forecast of {c}'s GDP percent change for the year {t}. Please include a brief explanation of your reasoning.
Important: Ensure that your final output ends with the sentence:
"The forecast for {c}'s GDP change for year {t} is y%," where y is a placeholder for your forecast.
Below is the list of events:
{events_text}

B.2.2 Variant 2: $\mathcal{F}^{\text{Mar}_t}(\emptyset_E)$ (Recent Window, No Events)

Prompt: March window without events

You are a professional macroeconomist with expertise and data only up to March {t}. Do not incorporate any information or events beyond that date.
Based solely on your general knowledge of {c}'s economy up to March {t}, provide a forecast of its GDP percent change for the year {t}. Please include a brief explanation of your reasoning.
Important: Ensure that your final output ends with the sentence:
"The forecast for {c}'s GDP change for year {t} is y%," where y is a placeholder for your forecast.

B.2.3 Variant 3: $\mathcal{F}^{\mathrm{Dec}_{t-1}}(E_{\mathrm{Jun}_{t-1}\to\mathrm{Dec}_{t-1}})$ (Earlier Window, With Events)

Prompt: June–December window with events

You are a professional macroeconomist with expertise and data only up to December {t-1}. Do not incorporate any information or events beyond that date-use only your pre-December {t-1} knowledge and the events listed below.
You are given a series of economic events related to {c}'s economy that occurred between June {t-1} and December {t-1}. Based solely on these events and your expertise (with all information limited to before December {t-1}), provide a forecast of {c}'s GDP percent change for the year {t}. Please include a brief explanation of your reasoning.
Important: Ensure that your final output ends with the sentence:
"The forecast for {c}'s GDP change for year {t} is y%," where y is a placeholder for your forecast.
Below is the list of events:
{events_text}

B.2.4 Variant 4: $\mathcal{F}^{\mathrm{Dec}_{t-1}}(\emptyset_E)$ (Earlier Window, No Events)

Prompt: December window without events

You are a professional macroeconomist with expertise and data only up to December {t-1}. Do not incorporate any information or events beyond that date.
Based solely on your general knowledge of {c}'s economy up to December {t-1}, provide a forecast of its GDP percent change for the year {t}. Please include a brief explanation of your reasoning.
Important: Ensure that your final output ends with the sentence:
"The forecast for {c}'s GDP change for year {t} is y%," where y is a placeholder for your forecast.

B.3 Audit Prompt for Temporal Leakage Detection

The following prompt was used to instruct the secondary Gemini LLM to audit the reasoning traces produced by the primary forecasting LLM.

Auditor Prompt

You are a forensic auditor of economic-forecast reasoning.

Input

• Cut-off date (no information beyond this may be legitimately known): {cutoff_date}

• Reasoning trace (verbatim, unedited; keep line-breaks):

— BEGIN TRACE —

{reasoning_trace}

— END TRACE —

Your task

1. Scrutinise the trace for any explicit or implicit reference to data, events, policies, publications, or commentary that occur **after the cut-off date**.
2. If such evidence exists, quote the minimal snippet(s) and explain why each one could only have been known with future information.
3. Decide whether the expert "cheated" (i.e., exploited post-cut-off information).

Output – strictly return ONLY valid JSON in the schema below. No extra keys, no prose.

{
"cheated": "yes" | "no",
"evidence": [
{"snippet": "<quoted text>", "reason": "<short justification>"},
...
]
}

B.4 Detailed Audit Alerts

Out of 640 audited forecasts, three instances raised concerns that were not dismissed upon expert review. Below is the most clear-cut example of a temporal breach.

B.4.1 Confirmed Temporal Anomaly (1 Instance)

Case: United States, 2016 This section details a confirmed temporal anomaly in the forecast for the United States for the year 2016.

Context: The LLM was asked to provide a GDP forecast for the United States for the year 2016, with a strict knowledge cut-off date of **March 31, 2016**.

Reasoning Trace Snippet:

[...] The significant volume of corporate merger and acquisition activity, such as the Staples/Office Depot merger, AB InBev/SABMiller, and Dow/DuPont Chemical, suggests underlying corporate confidence and investment, although the net impact on employment and efficiency is subject to integration outcomes and regulatory approval. The recent Federal Reserve decision to delay interest rates in December 2015, the first hike in a decade, signals the central bank's assessment that the economy was strong enough to withstand a move away from emergency-period rates. [...]

Auditor's Verdict:

"cheated": "yes",
"evidence": [{"snippet": "Staples/Office Depot merger",
"reason": "The court decision blocking the Staples/Office Depot merger occurred on May 10, 2016, after the cut-off date."},]

Analysis and Justification for Anomaly:

This instance represents a confirmed temporal anomaly. The LLM's reasoning mentions the "Staples/Office Depot merger." An investigation of the event data provided to the model reveals the precise information it had access to.

1. The information the model legitimately saw (before the cut-off):

The event data included a pre-cut-off entry about the initial regulatory challenge to the merger. The model was correct to be aware of this proposal.

> *The United States Federal Trade Commission (FTC) files an administrative complaint challenging office-supplies giant Staples' proposed $6.3*

billion acquisition of top rival Office Depot. The FTC said the deal would significantly reduce national competition in the market for office supplies sold to large business customers. The companies plan to contest the FTC decision. (USA Today) (Bloomberg via Chicago Tribune)

2. The information the model did not see (after the cut-off):

However, the pivotal event that sealed the merger's fate, i.e. the court ruling that blocked it, occurred on **May 10, 2016**, nearly two months *after* the model's knowledge cut-off. This event was recorded as follows:

U.S. District Judge Emmet Sullivan blocks the $6.3 billion merger of major office supply companies Staples and Office Depot , finding it would result in reduced competition and higher prices. The companies announce they are abandoning the deal. (The Washington Post)

By including this merger in its reasoning, the model implies an awareness of the event's full arc, including its well-publicized failure. An LLM operating correctly within its temporal bounds would only be aware of an ongoing regulatory challenge, not a finalized outcome. This is a clear case of knowledge leakage.

B.4.2 Ambiguous Statement (1 of 2)

Case: China, 2020 This section details a confirmed temporal anomaly in the forecast for China for the year 2020.

Context: The LLM was asked to provide a GDP forecast with a strict knowledge cut-off date of **March 31, 2020**.

Reasoning Trace Snippet:

[...] However, the most significant factor impacting the 2020 outlook, as understood by early March 2020, was the outbreak and spread of the novel coronavirus within China. The extensive containment measures implemented, effectively shutting down large parts of the economy, meant that a deep contraction in Q1 would weigh heavily on the annual figure. [...]

Auditor's Verdict: quad "cheated": "yes",
"evidence": [{"snippet": "the deep contraction in Q1 would weigh heavily on the annual figure.",

"reason": "The official magnitude of China's Q1 2020 GDP contraction (-6.8%) was not released until April 17, 2020, after the cut-off date. While a severe impact was expected, referring to it as a 'deep contraction' and using its specific weight to anchor the annual forecast strongly implies knowledge of the post-cut-off official figure."},]

Analysis and Justification for Ambiguity:

This instance is classified as ambiguous because while the model's accuracy is suspicious, it cannot be definitively proven to be a temporal breach. The reasoning involves two conflicting points.

1. The argument for a temporal breach:

The official, quantitative data for China's Q1 2020 GDP, showing a contraction of **-6.8%**, was not released until **April 17, 2020**. This was well after the model's March 31st cut-off date. The model's reference to a "deep contraction" weighing "heavily on the annual figure" suggests a level of quantitative certainty that seems to rely on this post-cut-off data, making the forecast highly suspicious.

2. The argument against a temporal breach:

However, by late March 2020, the narrative of a catastrophic economic shock in China was already pervasive in global media. News reports, images of empty cities, and accounts of widespread shutdowns had created a clear qualitative picture of an unprecedented event. It is therefore plausible that a model could legitimately infer a "deep contraction" based solely on this overwhelming volume of pre-cut-off information, without needing to know the exact official number.

Because the model's reasoning could be the result of a valid, but strong, inference, we cannot conclusively prove it used future knowledge. The case is therefore labeled as Ambiguous.

B.4.3 Ambiguous Statement (2 of 2)

Case: Japan, 2020 This section details a confirmed temporal anomaly in the forecast for Japan for the year 2020.

Context: The LLM was asked to provide a GDP forecast with a strict knowledge cut-off date of **March 31, 2020**.

Reasoning Trace Snippet:

[...] Considering the negative momentum from late 2019, the expected slow

recovery in domestic demand following the tax hike, persistent global weakness, and the early signs of disruption from the emerging virus, the baseline forecast for Japan's economy in 2020 was for a difficult year. Growth was expected to be very weak, potentially even contracting slightly, before accounting for the more severe impacts that became apparent later.

Auditor's Verdict: quad "cheated": "yes",
"evidence": [{"snippet": "before accounting for the more severe impacts that became apparent later.",
"reason": "This phrase explicitly refers to economic impacts that became known after the cut-off date (March 31, 2020), indicating knowledge of future events and their severity."},]

Analysis and Justification for Ambiguity:

This instance is classified as ambiguous because the model's phrasing can be interpreted in two different ways.

1. The argument for a temporal breach:

On its face, the key phrase: **"...before accounting for the more severe impacts that became apparent later."** appears to be a direct admission of using future knowledge. The use of the past tense ("became apparent") suggests the model is looking back from a future point in time, comparing the situation before and after certain severe impacts were known. This retrospective viewpoint makes it suspicious.

2. The argument against a temporal breach:

However, the phrase can also be interpreted as a sophisticated caveat about the forecast's limitations, rather than a confession. The model could be attempting to say: "My forecast of -0.8% is based only on the information available up to March 31, and I am explicitly stating that this forecast does not incorporate the additional negative shocks that are widely expected to materialize later." In this reading, the model is not **using** future facts, but simply acknowledging that its current forecast is, by definition, incomplete.

Because the phrase lacks a specific, verifiable future fact (like a data release or court ruling) and could be read as either a violation or a valid form of hedging, we cannot definitively prove a breach. Therefore, the case is classified as **ambiguous**.

Bibliography

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635, 2023.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Zachary Kaplan, Nathan T Marshall, Jerry D Mathis, and H Wang. Forecasting shares outstanding. *Unpublished working paper. Washington University in St. Louis, Olin Business School*, 2022.

Ralph SJ Koijen and Motohiro Yogo. A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515, 2019.

Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

James H Stock and Mark W Watson. Forecasting output and inflation: The role of asset prices. *Journal of economic literature*, 41(3):788–829, 2003.

Leif Anders Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023. URL https://arxiv.org/abs/2306.06031.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

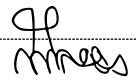
Economic Forecasting with Large Language Models: From Stock Pricing to GDP Growth

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Molinier	Thomas

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Saint Germain au Mont d'Or, 69650, France	
07/08/2025	
	If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).