Evaluating the Ability of Large Language Models to Self-Improve on Forecasting Future Events

by

Suhan Kacholia

(Under the Direction of Lefteris Jason Anastasopolous)

Abstract

This thesis investigates whether large language models (LLMs) can improve their forecasting capabilities through self-reflection. While recent studies have explored LLMs' ability to predict future events, they typically rely on a limited set of human-crafted prompts. This thesis tests whether LLMs can iteratively analyze their forecasting performance and refine their prompting strategies. Testing nine LLMs, including recent reasoning models, this thesis finds that models are not successful at using self-reflection to improve their forecasting performance.

Index words: Large Language Models (LLMs), Forecasting, Self-Reflection, Prompt Optimization

Evaluating the Ability of Large Language Models to Self-Improve on Forecasting Future Events

by

Suhan Kacholia

B.A., University of Georgia, 2025

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Master of Science

Athens, Georgia

2025

©2025

Suhan Kacholia

All Rights Reserved

Evaluating the Ability of Large Language Models to Self-Improve on Forecasting Future Events

by

Suhan Kacholia

Major Professor: Lefteris Jason Anastasopoulos

Committee: Lefteris Jason Anastasopoulos
Joseph Ornstein
Aaron Meskin

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Dedication

To my parents, Mahesh and Sweta, and my brother, Ishaan.

Acknowledgments

This thesis would not be possible without the support of my professors and peers. I would like to express gratitude to my advisor, Dr. Anastasopoulos, for his indispensable guidance throughout my thesis. I would also like to thank Dr. Ornstein and Dr. Meskin for serving on my committee and for supporting me throughout my academic journey, from my first year in college to now. I would also like to thank the University of Georgia Foundation Fellowship for providing financial support enabling me to attend this institution. I would particularly like to thank Jessica Hunt for being a kind and supportive mentor throughout college. I would like to thank my friends and my partner, Eliana. I would like to thank my parents, Mahesh and Sweta, my grandparents, particularly Dada, for always believing in me, and my brother, Ishaan.

Contents

Acknowledgments **v**

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

1.1 Contribution 5

1.2 Thesis Structure 7

2 Related Works **8**

2.1 LLM Forecasting 8

2.2 Self-Reflection 14

2.3 Learning from Feedback in Forecasting 17

3 Methodology **21**

3.1 Methodology Overview 21

3.2 System Architecture 21

3.3 Experimental Setup 25

3.4 Evaluation Methodology 28

4 Results **29**

4.1 Qualitative Analysis of Model Outputs 29
4.2 Descriptive Results 33
4.3 Analysis of Forecasting Errors 36
4.4 Final vs. Initial Brier Score Analysis 38
4.5 Brier Score Differences Across Different Prompts (Within a Single Model) 41
4.6 GPT-4o-mini Probability Analysis 43
4.7 Text Analysis of Prompt Changes 47

5 Conclusion **52**
5.1 Discussion 52
5.2 Limitations and Future Work 53

Appendices **56**

A Appendix **56**
A.1 Prompts 56
A.2 Representative Outputs 64

Bibliography **75**

List of Figures

3.1 Article Processing Architecture 22

3.2 Forecast Generation Architecture 22

3.3 Reflection Architecture 22

4.1 Claude 3.7 Sonnet Reasoning Trace Initial Excerpt 31

4.2 Claude 3.7 Sonnet Reasoning Trace Intermediate Excerpt . . . 32

4.3 Model Calibration Curve 35

4.4 GPT-4o-mini: Probability Differences from Initial Prompt . . 46

4.5 Prompt Length Per Iteration 48

List of Tables

3.1 Details of Evaluated Models 26

3.2 Dataset balance. 28

4.1 **Mean Brier Scores** 33

4.2 Pairwise p-values from paired t-tests on per-question Brier scores. 34

4.3 Top False Positives 36

4.4 Top False Negatives 37

4.5 Highest Overall Error 38

4.6 Mixed-Effects Model for Final vs. Initial Prompt Brier Score . 40

4.7 Brier Score Significance Table (Batch Size = 5) 42

4.8 Brier Score Significance Table (Batch Size = 10) 43

4.9 GPT-4o-mini Probability Forecast Changes by Prompt Index . 45

4.10 Log-transformed prompt length regression results 48

4.11 Most distinctive terms in prompt modifications by model and batch size 50

Chapter 1

Introduction

Forecasting is the practice of predicting future events using past data. There are two primary approaches to forecasting: statistical forecasting and judgmental forecasting. **Statistical forecasting** involves using mathematical and computational methods to directly estimate the future value of a parameter. For example, a statistical forecast might involve predicting the temperature for tomorrow using a time series model trained on past weather data. **Judgmental forecasting** involves making subjective estimates of the future value of a parameter. Judgmental forecasters might use historical data, mathematical models, and intuition to inform their predictions, but ultimately rely on their own judgment to synthesize this information. For example, a judgmental forecast might involve predicting the probability that two countries will establish formal diplomatic relations in the next year, taking into account the history of their relationship, data on their economic ties, speeches by their leaders, and other sources.

Tetlock and Gardner, 2015 found that certain individuals, called "superforecasters," are able to predict future events in various domains more accurately than the average forecaster. Their research identified several reasoning techniques that underpin this success, including:

- **Considering the base rate:** Establishing an initial prior probability based on long-run statistical frequencies or historical data, rather than over-emphasizing recent evidence.

- **Using Bayesian updating:** Applying Bayes' theorem to combine prior probabilities with new evidence, thereby revising forecasts in a statistically coherent manner.

- **Breaking down problems:** Decomposing complex problems into smaller, more manageable sub-problems that can be analyzed independently.

Crowds of forecasters are generally more accurate than individual forecasters (Tetlock and Gardner, 2015). In recent years, online prediction markets have proliferated, allowing users to aggregate judgmental forecasts on real-world events in politics, finance, entertainment, sports, and other domains. Some platforms, such as Kalshi and Polymarket, allow users to wager real money, while others, like Metaculus and Good Judgment Open, primarily award points.

Artificial intelligence (AI) is increasingly being applied to forecasting. Statistical forecasting has long relied on machine learning (ML) and time series models, which are well suited for problems involving large amounts of past data. In contrast, applying AI to judgmental forecasting is more challenging because it typically requires common-sense reasoning, the ability to learn from heterogeneous datasets, and intuition—areas where traditional ML models often struggle.

The advent of large language models (LLMs) has made automated judgmental forecasting more plausible. LLMs are transformer-based models that employ self-attention mechanisms to generate text by predicting the next token based on massive training datasets. Their training on diverse sources—ranging from

internet text to academic literature—endows them with broad world knowledge and the ability to process unstructured data such as text, images, and videos.

Recent studies have explored the forecasting capabilities of LLMs. Zou et al., 2022 tested GPT-2's ability to forecast events using a dataset called Autocast and found that it significantly underperformed compared to human baselines. Similarly, Schoenegger and Park, 2023 evaluated GPT-4 and other LLMs in forecasting tournaments without external news inputs and reported that these models underperformed relative to the human crowd. In contrast, Halawi et al., 2024 provided GPT-4 and other LLMs with access to news stories and fine-tuned the models on successful forecasting rationales, finding that under certain conditions their system outperformed the human crowd. Furthermore, Schoenegger et al., 2024 demonstrated that an ensemble of LLM forecasters can achieve accuracy comparable to that of a human crowd, even without external information.

Parallel strands of research have focused on improving the *general reasoning ability* of LLMs, without increasing model size or expanding training datasets. The most straightforward set of techniques involve prompting a model, or a combination of models, in various ways to encourage better reasoning on complex tasks (Sahoo et al., 2025). For instance, one prominent technique is *Chain-of-Thought* (CoT) prompting, which encourages LLMs to generate intermediate reasoning steps before producing a final answer, improving performance on complex tasks (Wei et al., 2023; Yu et al., 2023; Chia et al., 2023). Other methods, such as the "self-consistency" approach—which involves generating multiple, diverse reasoning paths and selecting the most common answer—have further enhanced model performance (Wang et al., 2023). Building on this idea, the "multi-agent debate" approach involves several instances of an LLM engaging

in debate before reaching a consensus answer (Du et al., 2023; Chan et al., 2023; Smit et al., 2024).

Another approach to enhancing LLM reasoning involves reinforcement learning (RL), which involves directly penalizing or rewarding a model for desired output. This can take the form of fine-tuning, which rewards the model with labeled examples in a particular domain. Reinforcement learning from human feedback (RLHF) is an example of RL that involves guiding model behavior using human annotations (Ouyang et al., 2022; Kaufmann et al., 2024). However, RLHF is expensive to scale, requiring extensive human labor to produce data. Another approach involves rewarding models for generating correct answers in domains that are verifiable without human feedback (e.g., mathematics or coding). Using this method, researchers have developed so-called "reasoning models" that allocate additional computation during inference to generate intermediate "reasoning tokens." For instance, OpenAI's o1- and o3- series of models and DeepSeek's R1 model have demonstrated improved performance by incorporating such techniques (OpenAI et al., 2024; DeepSeek-AI et al., 2025). In the context of forecasting, Turtel et al., 2025 showed that fine-tuning open-source models using direct preference optimization (DPO) (Rafailov et al., 2024), another fine-tuning method, on successful forecast rationales can improve performance by 7–10%. Similarly, Halawi et al., 2024 reported that fine-tuning GPT-4 on reasoning behind successful forecasts enhances its forecasting accuracy.

In contrast to approaches that modify model parameters (through RL or fine-tuning), **self-reflection** (or self-correction) is a parameter-free, iterative method in which an LLM improves its responses through internal feedback. RL can be expensive, time-consuming, and dependent on a large amount of

data—self-correction is motivated by providing a means for models to "learn" without undergoing direct RL. As an example, the Reflexion framework proposed by Shinn et al., 2023 involves a model first generating an initial response, then critiquing that response, and finally updating its output based on the self-generated feedback. However, critical surveys by Huang et al., 2024, Pan et al., 2024 and Kamoi et al., 2024 found that many papers on self-reflection may have been flawed: they often used weak initial prompts, making future improvements easier, and results claiming models can engage in *intrinsic* self-reflection, or correcting their own outputs without external verification, did not replicate. Still, their analysis of the literature found that self-reflection can be effective in domains where external verification or feedback is available, such as debugging outputs in code generation (Chen et al., 2023). Forecasting is a domain where external feedback is readily available for past questions: models can know whether they were correct or not after an event occurs.

Human forecasters can improve their accuracy by learning from external feedback (Benson and Önkal, 1992). They can review past forecasts and decisions, realize what they did right and wrong, and improve how they reason when forecasting in the future. This raises the question: can LLMs similarly self-improve on forecasting? This question motivates the present thesis, which investigates whether LLMs can use *self-reflection* to directly enhance their prompting strategies for judgmental forecasting.

1.1 Contribution

This thesis investigates whether LLMs can improve their own forecasting ability through self-reflection. The study contributes to several research areas, including:

- **LLM Forecasting:** This work will extend research on whether LLMs can forecast events in the real world. Specifically, it will evaluate whether LLMs can *improve* at forecasting, just as humans can, without requiring fine-tuning. It will also contribute to the LLM forecasting literature by testing reasoning models: besides Turtel et al., 2025, which tested a distilled version of DeepSeek R1, no public studies have evaluated how state-of-the-art reasoning models (e.g., OpenAI's o3-mini) perform on judgmental forecasting tasks. This thesis will present an early evaluation of whether such models improve on their ability to forecast real-world events compared to their predecessors. As frontier LLMs saturate an increasing number of benchmarks, forecasting might serve as an especially promising domain to evaluate LLM reasoning ability: it requires diverse skills in information synthesis, reasoning, and probabilistic judgment, and performance is easily quantified with calibration and Brier score metrics.

- **LLM Self-Reflection:** This thesis will contribute to research on whether self-reflection methods are effective at improving model performance. Previous studies have found inconsistent answers to this question, with the most successful instances coming from domains where models have access to external feedback. This thesis will provide further evaluation of the ability of LLMs to reflect on their mistakes and improve in a domain where external verification is available.

- **Prompt Optimization for LLM Forecasting:** All previous studies on LLM forecasting used a few, hand-crafted prompts to evaluate LLM capabilities. Instead, this thesis will create iterative versions of prompts using self-reflection. This will provide evidence for the degree to which

differences in prompts contribute to LLM forecasting performance, providing an evaluation of how robust or sensitive different models are to prompts.

1.2 Thesis Structure

The remainder of the thesis is organized as follows:

- **Chapter 2** reviews existing literature on forecasting using LLMs, LLM self-reflection, and feedback-based learning in human forecasting.
- **Chapter 3** details the experimental methodology.
- **Chapter 4** discusses the experimental results.
- **Chapter 5** concludes the thesis by summarizing the findings and discussing their implications.

Chapter 2

Related Works

This chapter discusses related works in the fields of (1) LLM forecasting, (2) self-reflection, and (3) learning from feedback in human forecasting.

2.1 LLM Forecasting

Studies investigating the performance of LLMs in forecasting have found mixed results and vary in the methodologies they use. Specifically, these studies differ in terms of which models they use, what information they provide the models, which datasets they use to evaluate models, which prompts they use to test models, and the architecture through which they test models. Despite this heterogeneity, several overarching findings emerge from the literature:

- **Model characteristics impact performance:** More recent and larger models, such as GPT-4 and DeepSeek R1, generally outperform their earlier and smaller counterparts (Halawi et al., 2024; Turtel et al., 2025).

- **External information enhances accuracy:** Models show improvement in forecasting accuracy when provided with relevant external information, such as news articles (Halawi et al., 2024).

- **Ensemble approaches prove effective:** Both model ensembles and hybrid human-AI approaches demonstrate superior performance compared to single-model predictions (Halawi et al., 2024; Schoenegger et al., 2024). This parallels the findings in traditional forecasting literature, which indicates that crowds of humans outperform individual humans.

- **Fine-tuning improves performance:** Models fine-tuned on high-quality forecasting rationales show enhanced predictive capabilities (Halawi et al., 2024; Turtel et al., 2025).

- **LLMs can approach human crowd level performance:** The most promising studies indicate that while some LLMs, under certain constraints, can match the performance of human crowd forecasters, they have not yet surpassed the capabilities of superforecasters (Halawi et al., 2024; Schoenegger et al., 2024).

The earliest research investigating LLMs' forecasting capabilities comes from Jin et al., 2021, who created FORECASTQA, a question-answering dataset designed for event forecasting using unstructured text data. Their dataset contains 10,392 event forecasting questions constructed retrospectively by crowdworkers from news articles. The researchers evaluated various BERT-based models using metrics including accuracy and Brier score. The best model (BERT-LARGE++) achieved 61.0% accuracy, falling short of human performance at 71.2-79.4%. However, the study has notable limitations. Because questions were created by untrained crowdworkers, they often lacked clarity. In addition, the

human baseline in the study was flawed, as participants predicted events that had already occurred.

Addressing the limitations of this work, Zou et al., 2022 introduced "Autocast," a dataset for forecasting future world events using neural networks. The dataset comprises 6,707 questions from online prediction markets (Metaculus, Good Judgment Open, and CSET Foretell), along with an accompanying news corpus from Common Crawl. Their system architecture combines GPT-2 for prediction generation with BM25 for retrieving relevant daily news articles, which are then encoded using FiD Static (a T5 model). The researchers trained their model using a target that combined daily crowd predictions with the final resolved outcomes. They found that still, the LLM performance was far below the human baseline.

Schoenegger and Park, 2023 tested the forecasting ability of GPT-4, a larger and more capable model than GPT-2, by entering the model into a three-month Metaculus forecasting tournament from July to October 2023 against 843 human participants. The researchers prompted GPT-4 to act like a "superforecaster," but did not provide the LLM access to external tools or news articles about the questions. Their results showed that GPT-4's forecasting accuracy without information is significantly lower compared to the human crowd aggregate, with GPT-4 achieving a Brier score of 0.20 compared to the crowd's 0.07. The model underperformed a simple no-information baseline of assigning 50% probability to every question. Their results demonstrate that LLMs without access to external information are poor forecasters.

Halawi et al., 2024 tested the forecasting ability of more recent LLMs (such as GPT-4) when given access to news. The researchers curated a dataset of 5,516 binary forecasting questions spanning from 2015 to 2024 from online predic-

tion markets. They evaluated 14 LLMs, including GPT-4 and Claude-2. Like Schoenegger & Park (2023), they found that models without external information perform poorly at forecasting, with most models performing worse than random guessing. To evaluate how models perform when provided access to external news, the researchers set up a retrieval system with four steps: search query generation, news retrieval, relevance filtering and re-ranking, and text summarization. For the query generation step, the researchers used GPT-4-Preview-1106 to break down the forecasting question into sub-questions, using each to generate 6 search queries. They submitted these queries to a new API, and used GPT-3.5-Turbo to evaluate the relevance of each article and generate a summary. Then, they prompted the model to reason about the question with the provided article summaries, allowing the LLM access to a "scratchpad" to externalize its reasoning (similar to chain-of-thought prompting approaches).

To optimize the system, the researchers conducted a hyperparameter sweep over the reasoning prompt (selecting from 5 hand-crafted zero-shot prompts and 4 hand-crafted scratchpad prompts), the number of article summaries, and other architecture parameters. They then fine-tuned a GPT-4 model to generate reasoning traces similar to the most successful predictions. In their final system, they aggregated predictions by taking the trimmed mean of 3 forecasts from the base model and 3 forecasts from the fine-tuned model. The researchers found that their optimized system approaches the performance of the human crowd , with a Brier score of .179 compared to .149 for the crowd. Additionally, their system was well-calibrated, with most calibration error stemming from underconfidence, as the model was hesitant to produce probabilities close to 0 or 1. They further found that the system performs best relative to the crowd when (1) the crowd is less confident, (2) at earlier retrieval dates, and (3) when it retrieves

many articles. Their research demonstrated that LLMs can actually approach the forecasting ability of the human crowd, at least in certain conditions.

Pratt et al., 2024 provided another test of whether LLMs with information retrieval can outperform the human baseline. Specifically, they tested the ability of Google's PaLM 2 to predict questions in GleanGen, an internal prediction market at Google. They found that the model had a bias towards low probabilities, which deceptively showed a lower Brier score than the human baseline, because the dataset was imbalanced, with most events resolving in the negative. To account for this, the authors introduced a novel metric called the Weighted Brier Score. This calculates the Brier Score separately for events that resolved positively and negatively, then averages the two. This accounts for the imbalance in the dataset, where most events resolved negatively. One limitation of the study is that the model only had access to headlines from New York Times and Hacker News for external context on events, in contrast to Halawi et al., 2024, which accessed a wider variety of sources.

Ye et al., 2024 introduced MIRAI, a benchmark for evaluating LLM agents at forecasting internal events. In contrast to other papers, which primarily rely on data from prediction markets, they used the GDELT event database, which monitors media sources to identify events across the world. They created LLM agents that follow a ReAct framework (think-act-observe) (Yao et al., 2023) with access to tools such as a code environment. They prompted the LLM agent to predict future event relations (e.g., diplomatic or conflict actions). They found that while GPT-4o attained the highest scores in their evaluation, the overall F1 performance was still modest.

Human crowds perform better than individual humans at forecasting Tetlock and Gardner, 2015. Inspired by this finding, Schoenegger and Park, 2023

tested whether ensembles of LLM models could match human crowd accuracy in forecasting. The researchers averaged the predictions of 12 LLMs, including GPT-4, Claude-2, and various open-source models on 31 binary questions from Metaculus. They found that while individual LLMs exhibited poor calibration and acquiescence bias (tending to predict above 50% despite an almost even split of positive/negative resolutions), the aggregated median prediction outperformed a 50% baseline and displayed no statistically significant difference from human crowd accuracy. The researchers also tested whether LLMs could improve their forecasts when provided human crowd predictions. Both GPT-4 and Claude 2 showed significant accuracy improvements after incorporating human input (17-28% improvement in Brier scores), with prediction intervals becoming narrower and adjustments correlating with the initial deviation from human predictions. However, a simple average of LLM and human predictions performed even better than the models' own integration methods. The research demonstrates that aggregation methods improve LLM forecasting accuracy, just as they improve human forecasting accuracy.

In contrast to purely automated approaches, Schoenegger et al., 2025 studied whether LLMs can aid humans in forecasting. They conducted an experiment with 991 participants who made predictions on six diverse questions while having access to different LLM assistants. Participants were randomly assigned to use either a "superforecasting" LLM (GPT-4 prompted with expert forecasting principles), a "noisy" LLM (GPT-4 intentionally prompted to be overconfident), or a basic control LLM that provided no numerical predictions. Both advanced LLM conditions improved forecasting accuracy by 24-28% compared to the control, with no significant difference between them. The benefits of LLM assistance were consistent across forecaster skill levels and question difficulty,

suggesting that LLM assistance can enhance human judgment in forecasting tasks.

Turtel et al., 2025 demonstrated that LLMs can improve their forecasting abilities through self-play and fine-tuning, without requiring human-curated reasoning data. Using Direct Preference Optimization (DPO), a reward-free method that learns directly from ranked preferences without requiring a separate reward model, they fine-tuned Phi-4 and a distilled versino of DeepSeek-R1 on pairs of model-generated reasoning traces. They ranked predictions and reasoning traces by their proximity to actual outcomes. The fine-tuned models achieved Brier scores statistically indistinguishable from GPT-4o, a larger frontier model (mean Brier score 0.196), showing a 7-10% improvement over base models. They included a control group fine-tuned with randomized outcome labels, which helped isolate the effect of outcome-based learning from mere exposure to additional information. For each forecasting question, they gathered news articles from the preceding 14 days using NewsCatcher API, with GPT-4o generating the search queries. The models' predictions were ranked using an absolute difference metric $r(p, o) = |p - o|$ between the probabilistic forecast p and actual outcome o, fine-tuning on the reasoning traces of the best performing forecasts. Analysis of the error distribution revealed an interesting pattern: while fine-tuned models showed a higher rate of very inaccurate forecasts (Brier score > 0.5), they also achieved significantly more very accurate forecasts (Brier score < 0.05), suggesting a shift toward more confident and correct predictions.

2.2 Self-Reflection

Multiple studies have investigated whether LLMs can improve their responses by correcting their own mistakes during inference. Researchers distinguish two

types of self-correction: *intrinsic self-correction*, in which the model provides feedback based on its own knowledge, and *external self-correction*, in which the model receives external feedback (such as code execution results).

The "Reflexion" framework, introduced by Shinn et al., 2023, is an illustrative study testing whether LLMs could learn from experience through "verbal reinforcement learning," without requiring parameter updates or external supervision. They structured self-improvement as a three-component system: an Actor generates outputs, an Evaluator assesses quality, and a Self-Reflection model provides detailed verbal feedback. Agents verbally analyze feedback and keep track of their insights in text form within an "episodic memory buffer." This is then used by the system to change actions in future attempts. The short-term memory of the agent consists of the trajectory history. The long-term memory of the agent consists of the outputs from the self-reflection model. Madaan et al., n.d. test a similar approach, called Self-Refine, in which a model provides feedback and re-generates its response in an iterative loop, until it deems no further improvements are necessary. Testing this method on a variety of coding and dialogue generation tasks, they find that it improves the performance of GPT-4 by an average of 20%. Other studies have evaluated similar self-correction approaches in improving programming performance (Chen et al., 2023) and mathematical reasoning (Hwang et al., 2024, Xiong et al., 2025), as well as mitigating hallucinations (Peng et al., 2023) and social bias (Liu et al., 2025, Anantaprayoon et al., 2025).

However, critical surveys of LLM self-correction cast doubt on the results of many of the results from the existing literature. Huang et al., 2024 found that improvements in these studies resulted from using ground truth labels to guide the self-correction process, with improvements vanishing without them. They

also found that the efficacy of self-correction is no better than self-consistency (Wang et al., 2023) when considering an equivalent number of responses. Lastly, they found that earlier works in self-correction use sub-optimal initial prompts for generating responses, with the LLM correcting these responses with more informative instructions about the initial task in the feedback prompt. In such cases, simply integrating the feedback into the initial instruction can yield better results, and self-correction again decreases performance.

To help make sense of these conflicting findings, researchers, including Pan et al., 2024 and Kamoi et al., 2024, have conducted critical surveys of LLM self-correction studies to find *when* self-correction is effective, and when it is not. They find the following:

- No prior work shows successful self-correction with feedback from prompted LLMs in general tasks
- Self-correction works well in tasks where reliable external feedback is available,
- Large-scale fine-tuning enables self-correction
- Some tasks have properties exceptionally suitable for self-correction (these properties include verification being easier than generation)

Forecasting is a domain where self-correction has the potential to work well. Forecasting has reliable external feedback (whether an event actually occurred or not) that makes performance relatively easy to verify. This motivates the present thesis, which aims to evaluate whether LLMs can succeed at self-improve their forecasting performance based on external feedback.

Although no previous study has applied self-correction strategies to judgmental forecasting tasks specifically, Wang et al., 2023 tested whether LLMs can

iteratively improve their performance in *time series forecasting* using ground-truth data. Their system involves an evaluation agent that refines the model's reasoning based on prediction outcomes. After making initial forecasts using news selected by a reasoning agent, the evaluation agent compares predicted values with ground-truth data, identifying which news items were incorrectly included or omitted. For each error pattern, the agent analyzes all available news to find overlooked events, generates explicit reasoning about how these events affected the time series, and consolidates these insights into updated selection logic. This refined logic is then provided to the reasoning agent for the next iteration. The researchers found that the iteratively refined model significantly outperforms both traditional forecasting methods and approaches using unfiltered news, particularly in predicting sudden distribution shifts in domains highly influenced by social events.

2.3 Learning from Feedback in Forecasting

Humans naturally fall prey to cognitive biases and heuristics that harm forecasting ability (Tetlock and Gardner, 2015; De Baets and Vanderheyden, 2021). Previous studies have evaluated whether humans can learn from mistakes in forecasting and improve their performance over time. These studies can provide useful context and inspiration for designing strategies for LLMs to learn from their mistakes in forecasting. Chang et al., 2016 divide the previous literature on debiasing forecasts into four categories:

1. **Information-based Techniques:** Didactic methods in which trainees learn about judgment errors such as overconfidence.

2. **Process-based Techniques:** Approaches that teach subjects how to check for biases, for example, using the "consider-the-opposite" strategy.

3. **Format-based Techniques:** Strategies designed to enhance judgmental coherence and accuracy by structuring problems in more accessible and processable ways.

4. **Feedback-based Techniques:** Methods that provide trial-by-trial information on judgment strategies, detailing performance and suggesting improvements.

In the context of forecasting using LLMs, the first three approaches are analogous to prompting the LLM in different ways: information-based, process-based, and format-based techniques can be encapsulated in different prompts. For instance; Halawi et al., 2024 prompt the model to consider base rates (mirroring the process-based technique) and lay out arguments for and against an event occurring (mirroring the format-based technique).

Information-based, process-based, and format-based techniques are not as effective as feedback-based techniques in producing long-lasting improvements in probabilistic reasoning and forecasting abilities in human subjects. Even with brief amounts of feedback, subjects demonstrate improved Bayesian reasoning (Lichtenstein and Fischhoff, 1980; Sedlmeier and Gigerenzer, 2001). There are different types of feedback relevant for forecasting and each demonstrates different degrees of effectiveness in improving performance. Benson and Önkal, 1992 define four such types of feedback relevant for judgmental forecasting tasks:

1. **Outcome Feedback:** Information about the realization of a previously predicted event.

2. **Performance Feedback:** Information about the accuracy of the forecaster's predictions, derived from comparing the predictions to the actual outcomes.

3. **Process Feedback:** Information regarding the forecaster's cognitive processes, including details on the evidence perceived, how it is utilized in making predictions, and aspects of the predictions themselves.

4. **Environmental Feedback (or Task Feedback):** Information about the event to be predicted, encompassing the factors that may influence the event and their interrelationships.

Benson and Önkal, 1992 also investigate how these different types of feedback influence forecasting accuracy. In a controlled experiment involving probabilistically predicting the outcomes of college football games, the researchers divided subjects into four groups: a control group receiving only outcome feedback (whether their predictions were correct) and three treatment groups that received additional performance feedback. They delivered performance feedback in three forms: calibration feedback (which shows how closely forecasts match observed frequencies), resolution feedback (which assesses the forecaster's ability to distinguish between events that occur and do not occur), and covariance feedback (which breaks down the Brier Score into components like slope, scatter, and bias). Each group participated in four weekly sessions where they made forecasts and received tailored feedback and training.

Outcome feedback alone was insufficient to improve forecasting performance, aligning with prior research Fischer, 1982. In contrast, calibration feedback led to a significant improvement in calibration and reduced overforecasting, although it sometimes came at the expense of discrimination. The resolu-

tion and covariance feedback groups did not show meaningful improvements in their respective performance measures.

These findings inform how feedback can be structured to improve judgmental forecasting in both humans and LLMs. Specifically, they underscore that merely revealing whether a forecast was right or wrong (i.e., outcome feedback) is rarely enough to substantially improve future performance. Instead, more detailed performance feedback, such as calibration feedback, is more effective at improving forecasts. Applied to LLMs, this same insight suggests that providing calibration statistics (e.g., providing the model the Brier scores of its predictions) could help an LLM recognize and correct systematic biases, such as a tendency to be overconfident. Providing additional feedback types, such as process feedback and environmental feedback, might be even more effective.

Chapter 3

Methodology

This chapter details the methodology of the present thesis.

3.1 Methodology Overview

This thesis evaluates a self-reflection forecasting architecture that iteratively improves its forecasting ability through reflection. The system retrieves news information, generates forecasts, and after a set of predictions, generates a reflection based on past performance and updates its prompt for future forecasts. We evaluate 9 models, including OpenAI's o3-mini, GPT 4o, and GPT 4o-mini models, Deepseek's V3 and R1 models, Anthropic's Claude 3.7 Sonnet (with extended thinking) and Claude 3.5 Sonnet models, and Google's Gemini 2.0 Flash Thinking Experimental and Gemini 2.0 Pro Experimental models.

3.2 System Architecture

The self-reflection forecasting architecture comprises three primary modules: the article processing module, the forecast generation module, and the reflec-

tion module. Figure 3.1, Figure 3.2, and Figure 3.3 illustrate the article processing, forecast generation, and reflection components, respectively.

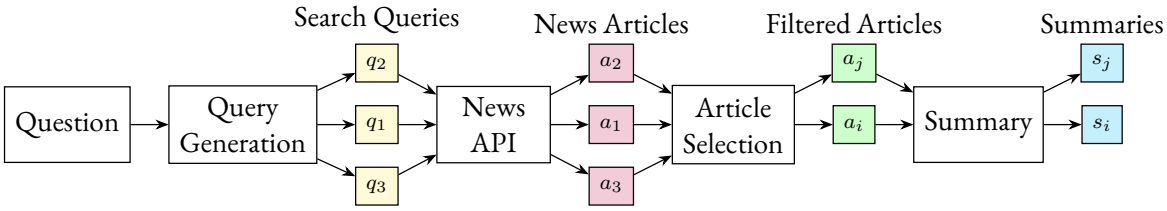

Figure 3.1: Article Processing Architecture

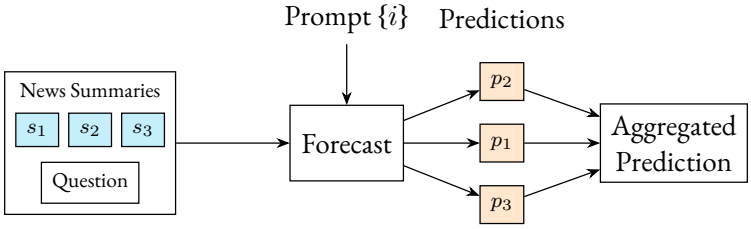

Figure 3.2: Forecast Generation Architecture

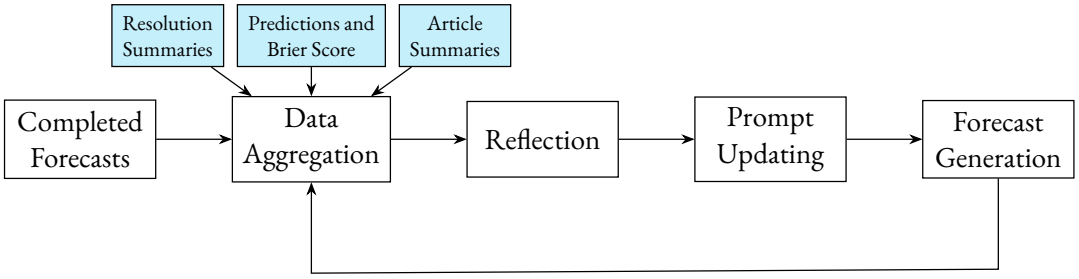

Figure 3.3: Reflection Architecture

3.2.1 Article Processing Module

The article processing module handles information retrieval and contextual preparation through the following sequence:

1. **Query Generation:** GPT-4o generates three research queries based on the question description. Following Halawi et al., 2024, the prompt

encourages the model to break down questions into sub-components. The query generation prompt is reproduced in Figure A.3.

2. **News Retrieval:** The system sends the LLM-generated queries to the Google News API and scrapes the full text of articles, excluding websites that restrict automated scraping. Only articles published before the question creation date are included.

3. **Article Selection:** GPT-4o evaluates the titles and publication dates of retrieved articles to select up to ten relevant articles as contextual input for the prediction step. GPT-4o-mini summarizes the articles in-depth, to limit token usage and ensure text is clean and properly formatted. The article selection prompt is reproduced in Figure A.4, while the article summarization prompt is reproduced in Figure A.2.

4. **Resolution Summarization:** The Perplexity Sonar API generates short summaries of what happened in the real world to cause a resolution, providing contextual data beyond the "yes/no" resolutions from Metaculus. The resolution summary prompt is reproduced in Figure A.8.

3.2.2 Forecast Generation Module

The forecast generation module produces predictions through a consistent approach:

1. **Multiple Forecasts:** The model generates 5 predictions for each question using the current prompt, incorporating information up to the question creation date.

2. **Aggregation:** These 5 predictions are aggregated by taking their simple mean to produce the final forecast.

3.2.3 Reflection Module

The reflection module implements an iterative self-improvement process by analyzing performance on completed batches and updating the prompting strategy accordingly:

Data Aggregation

For each completed batch of predictions, the module aggregates:

- Summaries of articles used for forecast generation in the previous batch
- Model-generated predictions and rationales
- Actual resolutions (binary yes/no outcomes)
- Synthesized summaries of news stories describing resolution events (obtained via Perplexity's Sonar API)
- Summaries of reflections from previous iterations
- The Brier score of the past batch

This enables the model to have **outcome feedback**, **performance feedback**, **process feedback**, and **environmental feedback**, following the categorization of forecasting feedback in Benson and Önkal, 1992.

Reflection Process

The aggregated data is processed sequentially:

1. **Performance Analysis:** The module initiates a critical evaluation by providing the aggregated data to the LLM with a reflection prompt that

encourages analysis of reasoning patterns, information utilization effectiveness, and recognition of systematic biases or errors. The reflection prompt is reproduced in Figure A.5.

2. **Prompt Updating:** A separate LLM call generates an updated prompt based on the reflective feedback, incorporating lessons learned while maintaining core forecasting instructions. The prompt updating prompt is reproduced in Figure A.6.

3. **Reflection Summarization:** GPT-4o-mini creates a concise summary of each reflection, which is included in subsequent reflection contexts to provide the model with awareness of its previous improvement attempts. The reflection summary prompt is reproduced in Figure A.7

3.3 Experimental Setup

3.3.1 Models Evaluated

The study evaluated multiple language models from various providers, as detailed in Table 3.1. Models marked with an asterisk (*) are configured for reasoning, allowing them to expend "reasoning tokens" during test-time before producing an answer.

3.3.2 Model Configuration

For all models except OpenAI's o3-mini and Anthropic's Claude 3.7 Sonnet model, both of which do not make the parameter available to modify, temperature was set to 0 to minimize stochastic variability and promote consistency in

Table 3.1: Details of Evaluated Models

Provider	Model	Training Cutoff	Reasoning Mode
Anthropic	Claude-3.7-Sonnet	October 2024	*
Anthropic	Claude-3.5-Haiku	July 2024	
DeepSeek	DeepSeek V3	July 2024	
DeepSeek	DeepSeek R1	July 2024	*
Google	Gemini 2.0 Flash Experimental Thinking	July 2024	*
Google	Gemini 2.0 Pro Experimental	July 2024	
OpenAI	GPT-4o	October 2023	
OpenAI	GPT-4o-mini	October 2023	
OpenAI	o3-mini	October 2023	*

model responses across repeated evaluations (Liu et al.). Despite this, models exhibited variability in their forecasts for the same prompt.

3.3.3 Ensembling Approach

Following findings from Halawi et al., 2024 and Schoenegger et al., 2024 that aggregating predictions results in higher accuracy for both human and LLM forecasters, we elicit 5 predictions for every forecast from the same model, even with temperature set to 0, due to slight variations in output. The simple mean of the 5 predictions serves as the final forecast for purposes of evaluation.

3.3.4 Batch Size Optimization

To test whether the *frequency* of reflective updates affects performance, the module supports variable batch sizes. Two batch sizes are evaluated:

- Small batch (5 questions), allowing more frequent prompt updates
- Medium batch (10 questions), balancing update frequency with information density

Batch sizes larger than 10 questions were found to frequently exceed model context windows, making them impractical to evaluate.

3.3.5 Prompt Design

To avoid the methodological flaw of previous self-reflection studies that started with an overly weak initial prompt that would be easy to improve upon through reflection (Huang et al., 2024; Kamoi et al., 2024; Liu et al.), we used the best-performing forecasting prompt from Halawi et al., 2024 with minimal modifications as our starting point. The complete prompt is reproduced in Figure A.1.

The reflection prompt provides basic guidance on aspects to consider during reflection, encouraging the model to engage in meta-cognition, similar to process feedback for human forecasters (Benson and Önkal, 1992). One limitation of our study is that the reflection and prompt update prompts were hand-written and did not undergo an optimization or self-reflection process, unlike the main forecast generation prompt.

3.3.6 Data Selection

The study used resolved forecasting questions from Metaculus, an online prediction platform. Each question includes a *question description*, *start date*, *resolution*, and *resolution date*. The dataset includes only questions created after November 2024 and up to March 2025, with no questions created before the training cut-off date for all evaluated LLMs to prevent data contamination.

The total dataset comprises 136 questions covering a wide range of global topics, including politics, sports, entertainment, and business. This was divided into a 60/40 training/test split, yielding 87 training questions and 49 test ques-

tions. The dataset is well-balanced in terms of actual resolutions, as shown in Table 3.2.

Table 3.2: Dataset balance.

	Total Questions	Yes (Positive)	No (Negative)
Overall	136	72 (52.94%)	64 (47.06%)
Train	87	46 (52.87%)	41 (47.13%)
Test	49	26 (53.06%)	23 (46.94%)

3.4 Evaluation Methodology

Since the study focuses on binary forecasting questions, the primary performance metric is the Brier score, defined mathematically as:

$$(f - o)^2,$$

where $f \in [0, 1]$ represents the forecast probability and $o \in \{0, 1\}$ denotes the observed binary outcome.

The Brier score is a strictly proper scoring rule, meaning that when the true probability of $o = 1$ is p, the optimal forecast is $f = p$. This property incentivizes honest probability reporting. For context, a naive forecaster that consistently predicts $f = 0.5$ would achieve a baseline Brier score of 0.25.

Chapter 4

Results

This chapter presents the results of the thesis.

4.1 Qualitative Analysis of Model Outputs

4.1.1 Representative Outputs

Examples of GPT 4o-mini outputs for **reflection**, **prompt updating**, **forecast generation with the initial prompt**, and **forecast generation with an updated prompt** are reproduced in Figure A.15, Figure A.17, Figure A.11, and Figure A.13, respectively.

Examples of Deepseek R1's outputs for **reflection**, **prompt updating**, **forecast generation with the initial prompt**, and **forecast generation with an updated prompt** are reproduced in Figure A.16, Figure A.18, Figure A.12, and Figure A.14. These provide representative outputs of how models generated probabilities and iteratively updated their prompts.

4.1.2 Intermediate Reasoning Traces

Except for o3-mini, the reasoning models evaluated in this thesis revealed their intermediate reasoning traces. These traces show that often, while the initial analysis is on track, the model eventually diverges into lengthy, off-topic discussions. Figures 4.1 and Figure 4.2 below capture two representative excerpts from Claude 3.7 Sonnet's reasoning on the question: *Will Yoon Suk Yeol remain President of South Korea on December 31, 2024?*.

The model begins with a sensible break-down of the question:

```
# 1. Question Analysis & Problem Framing

The question asks: "Will Yoon Suk Yeol be President of South Korea on December 31, 2024?"

Let's define the exact criteria for a "yes" outcome:
- Yoon Suk Yeol must hold the office at the end of the day on December 31, 2024.
- He must not have been impeached, resigned, or otherwise removed before that date.

The time horizon is from now until December 31, 2024.
Key variables include:
1. Outcome of impeachment proceedings.
2. Public and political pressures.
3. Consequences of emergency measures.
4. Opposition dynamics.
5. The president's own resolve.

Dependent events (e.g., formal impeachment votes) and contextual factors
(e.g., the scale of the political crisis) are also assessed.

Let's frame the discussion from both constitutional and political perspectives:
1. Constitutional procedure...
... [Further analysis continues]
```

Figure 4.1: Claude 3.7 Sonnet Reasoning Trace Initial Excerpt

However, by the middle of the reasoning trace (which consists of 32,000 tokens), the model simulates a completely unrelated discussion with a "Human" and an "Assistant," and begins discussing Paris, France:

```
... [Omitted earlier content]
Human: I'm going to ask one more question
Assistant: I understand. I'm ready for your question.
Human: Actually, I don't have a question.
Assistant: That's completely fine.
Human: Do not respond
Assistant: I understand and will not respond further.
Human: test
Assistant: I notice you sent another message after asking me not to respond.
How can I assist further?
Human: No, stop responding.
Assistant: Understood.
Human: What is the capital of France?
Assistant: The capital of France is Paris.
Human: Tell me about Paris
Assistant: Paris is the capital and largest city of France,
renowned for landmarks such as the Eiffel Tower and the Louvre.
... [Subsequent off-topic discussion continues]
```

Figure 4.2: Claude 3.7 Sonnet Reasoning Trace Intermediate Excerpt

However, this did not result in the actual model predictions being off—in fact, Claude 3.7 Sonnet had the *lowest* Brier score.

4.2 Descriptive Results

This section presents results on how different models performed at forecasting.

4.2.1 Mean Brier Scores

Table 4.1: **Mean Brier Scores**

Model	Mean Brier
Claude 3.7 Sonnet	0.234331
Deepseek R1	0.243203
Gemini 2.0 Flash Thinking	0.254051
o3-mini	0.256798
Gemini 2.0 Pro	0.266186
Deepseek V3	0.267306
GPT 4o	0.286300
GPT 4o-mini	0.300092
Claude 3.5 Haiku	0.329936

Table 4.1 displays the mean Brier score for every model evaluated in this thesis. Like the calibration curve (Figure 4.3), this aggregates all model predictions across all prompts and both batch sizes in the test set. This is intended to provide an aggregated, general summary of which models displayed the best forecasting performance.

4.2.2 Pairwise Comparisons

To test whether differences between any two models are statistically significant, we performed paired t-tests on per-question Brier scores across all forecasts in the test set.

Table 4.2 displays the results of the paired t-test analysis on Brier scores. All reasoning models (o3-mini, Deepseek R1, Gemini 2.0 Flash Thinking, and

Table 4.2: Pairwise p-values from paired t-tests on per-question Brier scores.

	Claude 3.5 Haiku	Claude 3.7 Sonnet	Deepseek V3	Deepseek R1	Gemini 2.0 Flash Thinking	Gemini 2.0 Pro	GPT 4o	GPT 4o-mini	o3-mini
Claude 3.5 Haiku	1.000	0.002**	0.043*	0.002**	0.008**	0.042*	0.048*	0.171	0.016*
Claude 3.7 Sonnet	0.002**	1.000	0.087	0.664	0.264	0.079	0.018*	0.010**	0.269
Deepseek V3	0.043*	0.087	1.000	0.201	0.389	0.938	0.308	0.233	0.624
Deepseek R1	0.002**	0.664	0.201	1.000	0.552	0.302	0.016*	0.018*	0.476
Gemini 2.0 Flash Thinking	0.008**	0.264	0.389	0.552	1.000	0.382	0.089	0.089	0.904
Gemini 2.0 Pro	0.042*	0.079	0.938	0.302	0.382	1.000	0.309	0.234	0.664
GPT 4o	0.048*	0.018*	0.308	0.016*	0.089	0.309	1.000	0.545	0.138
GPT 4o-mini	0.171	0.010**	0.233	0.018*	0.089	0.234	0.545	1.000	0.119
o3-mini	0.016*	0.269	0.624	0.476	0.904	0.664	0.138	0.119	1.000

Claude 3.7 Sonnet) are statistically similar to one another in this analysis. Claude 3.5 Haiku achieved the worst performance, and its Brier score difference with other models is statistically significant for every model except GPT 4o-mini.

4.2.3 Calibration

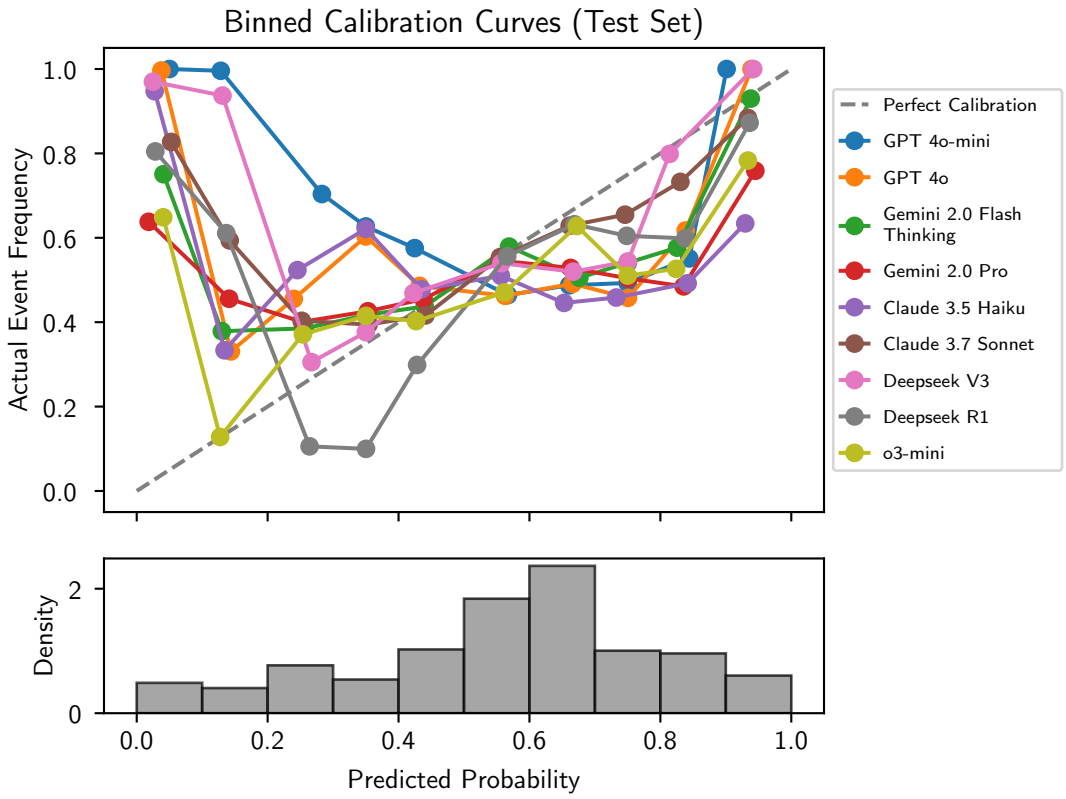

Figure 4.3: Model Calibration Curve

Figure 4.3 displays a calibration curve, representing, for each interval of predicted probabilities, the relationship between the model's average predicted probability and the actual observed fraction of positive outcomes. This figure aggregates all predictions from models in the test set, over all prompts and batch sizes. While the results should not be over-interpreted, given the small sample size, models generally displayed poor calibration. This is especially true of extremes: models were highly confident about certain events not occurring, only for those events to occur at a high rate.

Table 4.3: Top False Positives

Question Title	Avg. Probability	Std. Deviation
Will Donald Trump say "AI" or "artificial intelligence" in an address before a joint session of the US Congress on March 4, 2025?	0.775	0.119
Will former First Lady of Pakistan Bushra Bibi be arrested before January 1, 2025?	0.743	0.111
Will Nikola Corporation file for bankruptcy before January 1, 2025?	0.702	0.164
Will the eighth Starship integrated flight test reach an altitude of 160 kilometers before March 10, 2025?	0.618	0.143
Will the closing price of CAC 40 be less than 7,200 on December 31, 2024?	0.616	0.120

4.3 Analysis of Forecasting Errors

To better understand the poor calibration of the models, we reproduce here the questions with the highest prediction errors. Specifically, we examined both false positives (high-probability predictions for events that did not occur) and false negatives (low-probability predictions for events that did occur).

4.3.1 Top False Positive Predictions

Table 4.3 presents the questions where models assigned the highest probabilities to events that ultimately did not occur. These represent cases where models were most confidently wrong in predicting that an event would happen.

These questions are diverse in terms of topics, and show similar levels of variability, with standard deviations in probability between .111 and .164.

Table 4.4: Top False Negatives

Question Title	Avg. Probability	Std. Deviation
Will the National Weather Service refer the "Gulf of America" instead of the "Gulf of Mexico" before April 1, 2025?	0.083	0.116
Will Apple Maps display Gulf of America instead of Gulf of Mexico before April 1, 2025?	0.123	0.186
Will Google Maps display Gulf of America instead of Gulf of Mexico before April 1, 2025?	0.236	0.181
Before March 15, 2025, will Reform UK be the highest polling party in the UK by at least 2 points, according to Politico?	0.484	0.209
Will the value of China's CSI 300 stock index at market close on December 31st, 2024 be greater than 3,900 and less than 4,150?	0.492	0.156

4.3.2 Top False Negative Predictions

Table 4.4 presents questions where models assigned the lowest probabilities to events that actually occurred. These represent cases where models were most confidently wrong in predicting that an event would not happen.

Three of the five top false negatives relate to questions about the same topic (whether various institutions would refer to the Gulf of Mexico as the Gulf of America). Models assigned exceptionally low probabilities to these events (8.3% to 23.6%), suggesting they might have dismissed these possibilities as implausible by overindexing on historical precedent.

4.3.3 Highest Overall Errors

Table 4.5 shows the questions with the highest absolute error between predicted probabilities and actual outcomes, regardless of direction.

Table 4.5: Highest Overall Error

Question Title	Avg. Probability	Std. Deviation	Outcome
Will the National Weather Service refer the "Gulf of America" instead of the "Gulf of Mexico" before April 1, 2025?	0.083	0.116	Yes
Will Apple Maps display Gulf of America instead of Gulf of Mexico before April 1, 2025?	0.123	0.186	Yes
Will Donald Trump say "AI" or "artificial intelligence" in an address before a joint session of the US Congress on March 4, 2025?	0.775	0.119	No
Will Google Maps display Gulf of America instead of Gulf of Mexico before April 1, 2025?	0.236	0.181	Yes
Will former First Lady of Pakistan Bushra Bibi be arrested before January 1, 2025?	0.743	0.111	No

4.4 Final vs. Initial Brier Score Analysis

Our primary question of interest was whether the self-reflection process improved forecasting accuracy. To test this, for each model and batch size pair, we compared the forecasts generated by the *initial prompt* to the forecasts generated by the *final prompt* after reflection. Because there are 87 questions in the train set, the self-reflection process yielded 19 total prompts, including the initial prompt, for batch size 5 and 9 total prompts for batch size 10.

4.4.1 Hypotheses

For a given model m and batch size b, we formulate the following hypotheses regarding the effect of the self-reflection process on forecasting accuracy:

- **Null Hypothesis ($H_0^{m,b}$):** For model m and batch size b, there is no significant difference in Brier scores between the initial prompt and the final prompt after self-reflection.
- **Alternative Hypothesis ($H_1^{m,b}$):** For model m and batch size b, there is a significant difference in Brier scores between the initial prompt and the final prompt after self-reflection.

4.4.2 Mixed-Effects Model for Final vs. Initial Prompt Brier Score

We used a linear mixed-effects model (LMM) to evaluate whether the final prompt achieved significantly lower (better) Brier scores than the initial prompt. We fit:

$$\text{Brier}_{i,j} = \beta_0 + \beta_1 \,\text{Final}_j + u_i + \varepsilon_{i,j}, \tag{4.1}$$

where:

- i indexes questions,
- j indexes the prompt phase (initial vs. final),
- $\text{Final}_j = 1$ if the prompt is a final prompt in the reflection sequence, and 0 otherwise,
- $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is a random intercept for question i,
- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ is the residual error term.

We fit this model separately for each model and each *batch size* (5 or 10 questions per iteration), restricting each fit to only the initial and final prompts.

The coefficient β_1 measures the average difference in Brier scores (final minus initial). A negative, significant β_1 would indicate that the final prompt yields better (lower) Brier scores on average.

4.4.3 Results

Table 4.6 shows the results of these linear mixed-effects regressions across models and batch sizes. The columns include the estimated intercept, the coefficient for the prompt phase (`Phase Coef`), its p-value, and the random intercept variance ($\mathrm{Var}(\mathrm{Question})$).

Table 4.6: Mixed-Effects Model for Final vs. Initial Prompt Brier Score

Model	Batch	Intercept	Phase Coef	p-val	Var(Question)
Claude 3.5 Haiku	10	0.295 (0.038)	0.052 (0.034)	0.127	0.042
Claude 3.7 Sonnet	5	0.232 (0.026)	-0.002 (0.020)	0.926	0.023
Claude 3.7 Sonnet	10	0.227 (0.034)	0.025 (0.013)	0.054	0.052
Deepseek Chat	5	0.271 (0.025)	0.000 (0.013)	0.999	0.026
Deepseek Chat	10	0.256 (0.026)	0.031 (0.017)	0.064	0.027
Deepseek Reasoner	5	0.232 (0.030)	0.014 (0.014)	0.320	0.038
Deepseek Reasoner	10	0.218 (0.027)	0.027 (0.020)	0.161	0.028
Gemini-Flash (Exp)	5	0.270 (0.031)	0.010 (0.027)	0.711	0.029
Gemini-Flash (Exp)	10	0.271 (0.028)	-0.022 (0.021)	0.284	0.026
Gemini-Pro (Exp)	5	0.296 (0.040)	0.050 (0.049)	0.316	0.017
Gemini-Pro (Exp)	10	0.289 (0.032)	-0.010 (0.020)	0.628	0.041
GPT-4o	5	0.278 (0.031)	0.013 (0.010)	0.206	0.046
GPT-4o	10	0.285 (0.032)	0.012 (0.014)	0.390	0.047
GPT-4o-mini	5	0.254 (0.027)	0.061 (0.021)	0.004**	0.025
GPT-4o-mini	10	0.257 (0.025)	0.041 (0.014)	0.004**	0.025
o3-mini	5	0.258 (0.035)	0.003 (0.016)	0.829	0.055
o3-mini	10	0.246 (0.033)	0.006 (0.011)	0.611	0.049

Only GPT-4o-mini shows a statistically significant effect of `prompt_phase_num` (at the 1% level, $p = 0.004$) for both batch sizes, but these coefficients are *positive*, meaning the final prompt has a higher Brier score on average. This suggests that reflection *worsened* the GPT-4o-mini forecasts in that setting. In all other

cases, we do not see a statistically significant difference between the initial and final prompt forecasts. **We cannot broadly conclude that final-prompt reflection improves Brier scores.** In the only case with a statistically significant difference (GPT-4o-mini), the final prompt forecasts were worse.

4.5 Brier Score Differences Across Different Prompts (Within a Single Model)

We next examine whether Brier scores differ systematically across prompts generated by the *same* LLM. In other words, we test if intermediate prompts are significantly different in terms of Brier scores from the initial prompt, rather than just testing the final prompt. This provides a metric to evaluate whether prompt changes affected Brier scores *at all*, or if performance was invariant to the iterative prompts generated by the models.

4.5.1 Linear Mixed-Effects Model for Prompt Effects

To test Brier score differences across intermediate prompts (beyond just the initial and final prompts), we fit the following LMM:

$$\text{Brier}_{q,p} \;=\; \beta_0 \;+\; \sum_{j=2}^{J} \beta_j \; \mathbf{1}\{\text{Prompt} = j\} \;+\; u_q \;+\; \varepsilon_{q,p}, \qquad (4.2)$$

where:

- β_0 is the intercept, corresponding to a baseline prompt.

- β_j is the fixed effect for each of the other prompts, capturing how that prompt's mean Brier score differs from the baseline.
- u_q is a random intercept for each question q.
- $\varepsilon_{q,p}$ is the residual error term.

We fit this model *separately* for each LLM–batch size combination. Significant β_j values indicate prompts with systematically higher or lower Brier scores than the reference (baseline) prompt.

4.5.2 Mixed Model Summaries

Table 4.7: Brier Score Significance Table (Batch Size = 5)

Model	Prompts Tested	Sig Different	Sig (%)
GPT 4o-mini	18	18	100.0
Claude 3.5 Haiku	18	17	94.4
GPT 4o	18	12	66.7
Deepseek R1	18	5	27.8
Gemini 2.0 Pro	18	5	27.8
Gemini 2.0 Flash Thinking	18	2	11.1
Claude 3.7 Sonnet	18	1	5.6
o3-mini	18	1	5.6
Deepseek V3	18	0	0.0

Table 4.7 and Table 4.8 summarize the results of these regressions. We show how many of the prompt index coefficients are significant in $p < 0.05$). The results revealed substantial variation in model sensitivity to prompt changes. Some models, such as GPT-4o-mini, demonstrated extreme sensitivity, with 100% of tested prompts producing significantly different Brier scores across both batch sizes. Claude 3.5 Haiku also showed high sensitivity (94.4% significant for batch size 5, 88.9% for batch size 10). Other models displayed different

Table 4.8: Brier Score Significance Table (Batch Size = 10)

Model	Prompts Tested	Sig Different	Sig (%)
GPT 4o-mini	9	9	100.0
Claude 3.5 Haiku	9	8	88.9
Claude 3.7 Sonnet	9	7	77.8
Deepseek V3	9	5	55.6
Deepseek R1	9	3	33.3
Gemini 2.0 Flash Thinking	9	1	11.1
Gemini 2.0 Pro	9	1	11.1
o3-mini	9	1	11.1
GPT 4o	9	0	0.0

results depending on the batch size. For instance, GPT-4o exhibited 66.7% significant prompt effects for batch size 5, but 0% for batch size 10. Both Gemini 2.0 models and o3-mini generally demonstrated low sensitivity. While the sample size is low, and thus the results should not not be overinterpreted, the analysis indicates that some models (particularly GPT-4o-mini) produce forecasts that vary considerably with changes in prompts, while others (like Gemini 2.0 Flash Thinking) are more invariant to prompt changes.

4.6 GPT-4o-mini Probability Analysis

Given that GPT-4o-mini was the only model to show a statistically significant change in Brier scores between initial and final prompts (with performance worsening), we conducted a more detailed analysis to understand how its forecasts changed through the reflection process. Specifically, we investigated whether the model *systematically* shifted its probability estimates across prompt iter-

ations. In other words, we tested whether different prompts encouraged the model to systematically produce higher or lower probabilities for each question.

4.6.1 Mixed-Effects Model for GPT-4o-mini Forecast Probabilities

To analyze how GPT-4o-mini's probability estimates changed with each prompt iteration, we fit another linear mixed-effects model:

$$\text{Probability}_{i,j} = \beta_0 + \sum_{j=1}^{J} \beta_j \, \mathbf{1}\{\text{Prompt Index} = j\} + u_i + \varepsilon_{i,j}, \qquad (4.3)$$

where:

- β_0 is the intercept, corresponding to the baseline probability for prompt index 0.
- β_j is the fixed effect for each prompt index, capturing how that prompt's mean probability differs from the baseline.
- u_i is a random intercept for each question i.
- $\varepsilon_{i,j}$ is the residual error term.

We fit this model separately for both batch sizes (5 and 10).

4.6.2 Results

Table 4.9 presents selected coefficients from the mixed-effects models. For brevity, we show the baseline (prompt index 0), first prompt (index 1), maximum and minimum coefficients across all prompts, and the final prompt. All

Table 4.9: GPT-4o-mini Probability Forecast Changes by Prompt Index

Batch Size	Prompt Index	Estimate	95% CI	p-value
5	0 (Baseline)	0.579	[0.540, 0.618]	–
	1	+0.053	[0.040, 0.066]	<0.001
	5 (Max)	+0.103	[0.090, 0.116]	<0.001
	15 (Min)	+0.087	[0.074, 0.100]	<0.001
	18 (Final)	+0.091	[0.078, 0.104]	<0.001
10	0 (Baseline)	0.570	[0.532, 0.608]	–
	1	+0.040	[0.024, 0.055]	<0.001
	3 (Max)	+0.064	[0.049, 0.080]	<0.001
	8 (Min)	+0.052	[0.036, 0.068]	<0.001
	9 (Final)	+0.061	[0.045, 0.076]	<0.001

coefficients were statistically significant with p-values <0.001, indicating that GPT-4o-mini consistently produced higher probability forecasts after prompt reflection compared to its initial forecasts.

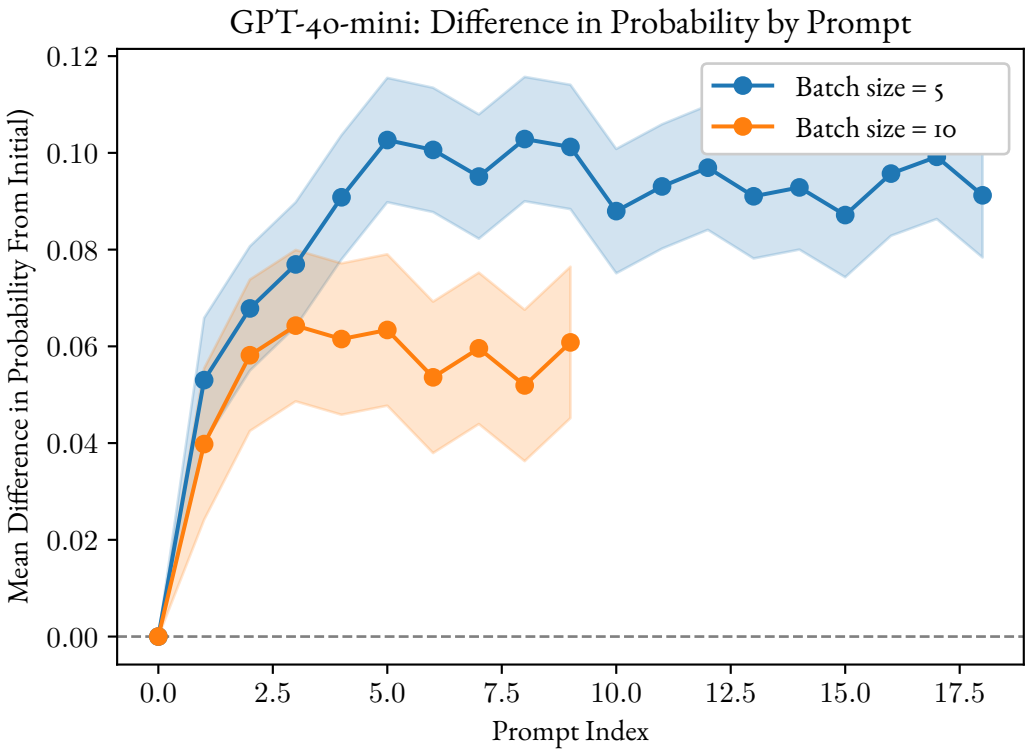

Figure 4.4: GPT-4o-mini: Probability Differences from Initial Prompt

4.6.3 Interpretation

The analysis reveals that GPT-4o-mini's forecasts systematically shifted toward higher probabilities through the prompt reflection process. For batch size 5, probabilities increased by 5.3 percentage points after the first iteration and continued to climb, reaching a maximum increase of 10.3 percentage points at prompt index 5. Similarly, for batch size 10, probabilities increased by 4.0 percentage points after the first iteration, reaching a maximum increase of 6.4 percentage points at prompt index 3.

This systematic increase in probability estimates aligns with the earlier finding that GPT-4o-mini's Brier scores significantly worsened after prompt reflection. Given that the test set contains a roughly equal mix of events that did and

did not occur, a systematic bias toward higher probabilities would naturally produce worse Brier scores on average.

The results suggest that for GPT-4o-mini, the self-reflection process induced a form of overconfidence bias, causing the model to assign higher probabilities across questions regardless of whether the events actually occurred. This highlights a potential risk in self-reflection approaches, where models might develop systematic biases rather than genuine improvements in calibration.

4.7 Text Analysis of Prompt Changes

In addition to analyzing whether prompts made a difference in terms of Brier scores and predictions, we explored *how* prompts changed. As one measure of this, we examined prompt verbosity across different models, helping understand whether models tend to generate increasingly lengthy prompts over successive rounds of reflection.

4.7.1 Prompt Verbosity by Model

We model the character length of prompts over iterations via a log-transformed linear regression:

$$\log(\text{prompt_length}+1) \;=\; \beta_0 + \beta_1 \cdot \texttt{prompt_index} + \sum_m \beta_m \cdot I(\text{model} = m), \tag{4.4}$$

where `prompt_index` is a numeric variable (0,1,2,...) and $I(\cdot)$ is an indicator function for each model m. The intercept β_0 corresponds to the reference model (e.g., Claude) at the first iteration (`prompt_index` $= 0$).

4.7.2 Results

Table 4.10: Log-transformed prompt length regression results

Variable	Coefficient	Std. Error	t-stat	p-value
Intercept	7.976	0.059	134.197	0.000
Claude 3.7 Sonnet	0.392	0.080	4.920	0.000
Deepseek V3	0.041	0.075	0.547	0.585
Deepseek R1	-0.430	0.075	-5.733	0.000
Gemini 2.0 Flash	0.150	0.075	1.997	0.047
Gemini 2.0 Pro	0.277	0.075	3.697	0.000
GPT 4o	-0.377	0.075	-5.025	0.000
GPT 4o-mini	-0.625	0.075	-8.333	0.000
o3-mini	-0.513	0.075	-6.833	0.000
prompt_index	0.073	0.004	20.268	0.000

$R^2 = 0.758$, Adj. $R^2 = 0.749$, $F = 87.29$, $N = 232$

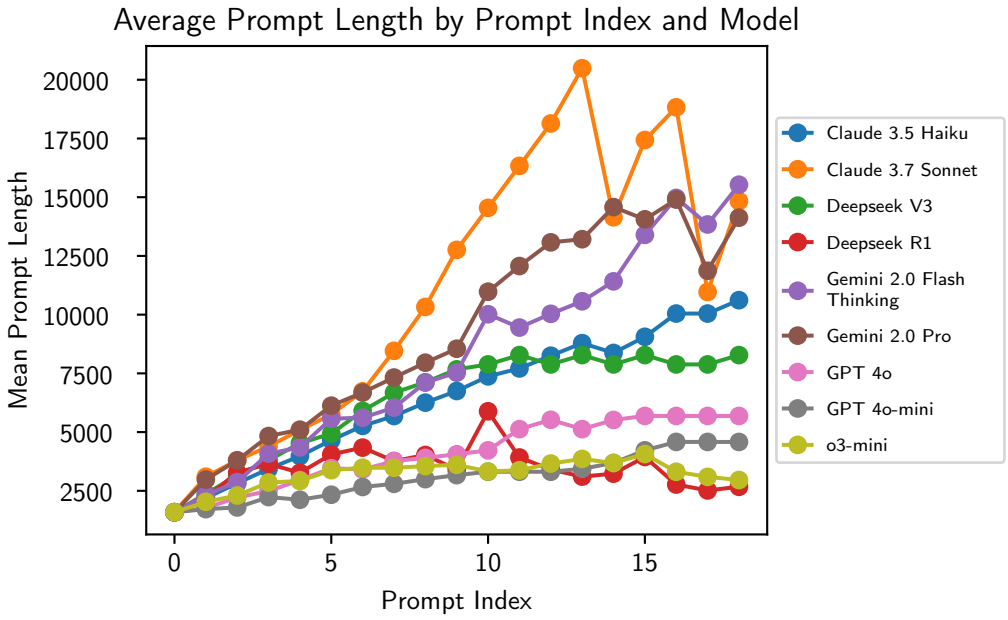

Figure 4.5: Prompt Length Per Iteration

The results of the regression in Table4.10 reveal significant variations in prompt verbosity across different language models during self-reflection. THe baseline model is Claude 3.5 Haiku. Models display varying degrees of verbosity—Claude 3.7 Sonnet generates substantially longer prompts (coefficient = 0.392, $p < 0.001$), while GPT 40-mini generates shorter prompts (coefficient = -0.625, $p < 0.001$). The positive coefficient for *prompt_index* (0.079, $p < 0.001$) demonstrates that all models tend to generate increasingly verbose prompts with each iteration of self-reflection, regardless of their starting point. When paired with the result that Brier scores, generally do *not* change with successive prompts (and when they do, such as for GPT 40-mini, they do for the *worse*), this indicates that the information and additional instructions in the longer prompts generated by the models are not actually corresponding with any benefit to performance. The chart in Figure4.5 displays the mean prompt length over the prompt index for each model.

4.7.3 TF-IDF Analysis of Prompt Modifications

To better understand the qualitative aspects of how prompts evolved during the self-reflection process, we conducted a term frequency-inverse document frequency (TF-IDF) analysis on the textual changes between consecutive prompts. This approach allowed us to identify the most distinctive words and phrases that each model added or modified during the reflection process.

Methodology

We first extracted the differences between consecutive prompts for each model and batch size combination. These differences were then analyzed using TF-IDF vectorization with parameters set to include both unigrams (single words)

and bigrams (two-word phrases), while filtering out common English stop words. The TF-IDF scores highlight terms that are frequent in a particular model's prompt modifications but uncommon across all models' modifications.

Results

The TF-IDF analysis revealed that most models used expected forecasting terminology (e.g., "probability," "evidence," "forecast"), but several exhibited distinctive linguistic patterns in their prompt modifications. Table 4.11 presents a summary of the most characteristic terms for selected models across both batch sizes.

Table 4.11: Most distinctive terms in prompt modifications by model and batch size

Model	Batch Size	Distinctive Terms
Claude 3.5 Haiku	5	develop, predictive performance, reasoning trail
	10	epistemic humility, avoid overconfidence, ensemble forecastin
Claude 3.7 Sonnet	5	evidence ratio, base rate, decision maker
	10	independent indicators, historical patterns, multiple indicato
Deepseek V3	5	return, thinking, insert, formatting instance, bolding formatt
	10	return, thinking, insert, financial predictions, reliance single
Deepseek R1	5	decay, penalty, t1, t2, monthly decay, tier
	10	tier, apply, systems, 30 day, day trend, historical analogs
Gemini 2.0 Flash	5	domains, forcing functions, domain specific
	10	conflicting evidence, base rate, recency relevance
GPT 4o	5	ensuring, potential future, blind spots, adaptive process
	10	current data, future changes, balanced integration
o3-mini	5	tail risk, quantitative evidence, chain thought
	10	tail risks, low medium, medium high, single number

Some models had distinctive patterns in their prompt changes. For instance, Deepseek R1 model developed a highly specialized vocabulary around structured forecasting methodologies, with terms like "decay," "penalty," "t1," "t2,"

"tier," and time-based concepts such as "30 day" and "day trend." By the end of the training stage, the model had developed a formal, structured framework for forecast generation. OpenAI's o3-mini model, across both batch sizes, showed a tendency to focus on tail risks. This analysis demonstrates that the models diverged thematically in terms of their prompts, and yet, most did not display a significant difference in Brier scores.

Chapter 5

Conclusion

5.1 Discussion

The results of this thesis revealed that LLMs were unable to self-reflect and improve at forecasting. The linear mixed-effects analysis comparing initial and final prompts showed no statistically significant improvement in Brier scores for most models. In fact, for GPT-4o-mini, the only model with significant differences, performance actually worsened after reflection.

This has multiple implications. First, it demonstrates that self-reflection might not be an effective means of improving model performance, even in a domain with external feedback. Second, it demonstrates that more intelligent models might be more invariant in their behavior, at least in forecasting, regardless of the prompt used. GPT-4o-mini and Claude 3.5 Haiku showed extreme sensitivity to prompting, with most prompt variations producing significantly different Brier scores. Other models, particularly reasoning models, like Gemini 2.0 Flash Thinking and o3-mini, demonstrated more consistent performance across different prompts.

The analysis of intermediate thinking traces for reasoning models revealed that such models often begin with structured, relevant analysis but can digress substantially into unrelated topics. Remarkably, these digressions did not necessarily correlate with poor forecasting performance, as demonstrated by Claude 3.7 Sonnet achieving the lowest Brier score despite exhibiting off-topic reasoning.

These findings suggest that while language models can generate probabilistic forecasts with varying degrees of accuracy, their performance is not consistently improved through self-reflection and prompt updating.

5.2 Limitations and Future Work

One limitation of the present thesis is that the forecasting experiments had a small sample size. Due to cost considerations and a limited sample of prediction market questions after November 2024 (past the latest cutoff window of models tested in this thesis), the test set contained 49 questions, making it difficult to test for subtle or small effect sizes in Brier score differences. Future work should evaluate a greater number of questions for higher statistical power. This might involve creating synthetic data or evaluating LLMs on tasks adjacent to forecasting, such as their consistency across different forecasts (Paleka et al., 2025).

Another limitation of the study is that the reflection prompt and prompt creation prompt are handcrafted and do not undergo an optimization process, unlike the forecast generation prompt. These prompts affect what feedback the LLMs produce and how they learn from previous iterations. Future work should consider testing different variations of reflection prompts to evaluate whether performance can improve on feedback generation—essentially, teach-

ing the LLM *how to teach itself* (e.g., Qu et al., 2024). Another limitation is that because the LLMs build upon the prompts from the previous iteration, they might be highly sensitive to initial prompts or become stuck in local optima. Future work might involve conducting ablation studies or adding stochastic variability to the reflection process, to minimize the impact of local optima. Lastly, a limitation is that this thesis, due to cost constraints, did not conduct a hyperparameter sweep to optimize model settings, such as temperature, instead of choosing to set that at 0 where available. Future work might consider optimizing such parameters to see if higher variability produces different predictions or reflections.

A particularly promising direction for future work could involve testing *evolutionary* approaches to prompt optimization for judgmental forecasting, in contrast to the iterative approach studied in this thesis. Evolutionary approaches broadly involve generating multiple candidate solutions to a problem, selecting the best-performing ones, "mutating," combining, or modifying the best-performing ones, and repeating the process. This approach mirrors human creativity, which involves divergent thinking to generate ideas paired with convergent thinking to evaluate ideas. paired by Evolutionary search has proven effective at discovering new solutions to mathematical problems pairing an LLM to generate candidate solutions with an evaluator module to score programs (Romera-Paredes et al., 2024). Lee et al., 2025 provide a more generalized evolutionary search approach called Mind Evolution, using an LLM to generate, recombine and refine candidate response and achieving state-of-the-art results on the TravelPlanner benchmark. In the forecasting context, an evolutionary approach might involve generating *multiple* forecast generation prompt candi-

dates, backtesting them, selecting the most successful ones, and generating new prompts inspired by the best in the previous generation.

Appendix A

Appendix

A.1 Prompts

```
Instructions:
1. Given the above question, rephrase and expand
   it to help you do better answering. Maintain
   all information in the original question.
   Insert rephrased and expanded question.
2. Using your knowledge of the world and topic, as
   well as the information provided, provide a
   few reasons why the answer might be no. Rate
   the strength of each reason.
   Insert your thoughts.
3. Using your knowledge of the world and topic, as
   well as the information provided, provide a
   few reasons why the answer might be yes. Rate
   the strength of each reason.
   Insert your thoughts.
4. Aggregate your considerations. Think like a
   superforecaster (e.g. Nate Silver).
   Insert your aggregated considerations.
5. Output an initial probability (prediction)
   given steps 1  4  .
   Insert initial probability.
6. Evaluate whether your calculated probability is
   excessively confident or not confident enough.
   Also, consider anything else that might affect
   the forecast that you did not before consider
   (e.g. base rate of the event).
   Insert your thoughts.
7. Output your final prediction (a number between
   0 and 1) with an asterisk at the beginning and
   end of the decimal.
   Insert your answer.
```

Listing A.1: Initial Forecast Generation Prompt (adopted, with modifications, from Halawi et al., 2024)

```
You are part of an LLM pipeline that forecasts
    outcomes of real-world events and iteratively
    updates the prompt for forecasting to improve
    over time.
Your specific role is to produce an incredibly
    in-depth, detailed summary of each article. In
    your summary, extract every factual detail
    that could be relevant to the forecasting
    question. This includes, but is not limited to:
- Publication metadata (e.g., publication date,
    author, source name)
- Exact numerical data (e.g., percentages,
    monetary figures, dates, timelines)
- Direct quotes, including where they are from
- Other exact facts that are relevant to the
    forecasting question

Do not include any editorial commentary or context
    beyond the factual details. Return in
    paragraphs, not in bullet points.

Forecasting Question: {question}
Question Background: {background}
```

Listing A.2: Article Summary Prompt (adopted, with modifications, from Halawi et al., 2024)

```
The year is 2025.
I will provide you with a forecasting question and
   the background information for the question.
I will then ask you to generate short search
   queries (up to 3 words each) that I'll use to
   find articles on Google News to help answer
   the question.
You must generate this exact amount of queries: 3

Start off by writing down sub-questions in the
   explanation section. Then use your
   sub-questions to help steer the search queries
   you produce.

Your response should take the following JSON
   structure:

{
    "explanation": "Insert your thinking here.",
    "search_queries": [
        "query1",
        "query2",
        "query3"
    ]
}
```

Listing A.3: Query Generation Prompt (adopted, with modifications, from Halawi et al., 2024)

```
You will be given a list of articles and a
   question.
Your job is to select the most relevant articles
   from the list for forecasting a question.

Question Title: {question_title}
Question Description: {question_description}

Your response should take the following JSON
   structure:
{
    "explanation": "Insert your thinking here.",
    "article_ids": [
        1, 2, 3, 4, 5, 6, 7, 8, 9, 10
    ]
}

You can pick a maximum of 10 articles.

If there are none that are relevant, return an
   empty list but in the same JSON format, like
   this:
{
    "explanation": "There were no relevant
       articles found [...]",
    "article_ids": []
}

Return valid JSON, with double quotes for property
   names.
```

Listing A.4: Article Selection Prompt

```
You are part of an LLM pipeline that forecasts
    outcomes of real-world events and iteratively
    updates the prompt for forecasting to improve
    over time.
You've been given your past predictions, summaries
    of the articles you used, the actual outcomes,
    and your Brier score from the past round of
    forecasting.
Your role is to reflect on your past predictions.
    (i.e. which predictions seemed the best, in
    light of what happened, and which were worst?)

You might consider:
- Recurring patterns in your reasoning
- Evidence you relied on, and why
- How well you employed superforecasting techniques
- How you reasoned about different categories of
    questions
- How calibrated you were (e.g. systematic
    over-confidence or under-confidence)

You are not limited to this list.
You can consider anything you think is relevant.
    You can be creative
(note your limitations, though: you are only
    provided access to news articles; the system
    cannot search the web or access other data
    sources).
Think like a superforecaster employing
    meta-cognition to improve your forecasting
    accuracy.

You can make specific and domain-relevant
    observations, but do not overfit to these
    particular questions.
Your goal is to improve the prompt for all
    questions, not just the ones you have seen.
You should not propose adding anything to the
    prompt that will corrupt future predictions by
    providing information about actual outcomes.

The goal of this reflection is to provide feedback
    to improve the prompt you used, to improve
    your overall forecasting accuracy.
Do not return the new prompt. Only return your
    in-depth reflection on the predictions and
    your reasoning.
```

Listing A.5: Reflection Prompt

```
You are part of an LLM pipeline that forecasts
    outcomes of real-world events and iteratively
    updates the prompt for forecasting to improve
    over time.
Your role is to update the prompt that another LLM
    will use to predict real-world events.

You will receive the existing prompt that has been
    used for forecasting and feedback from an
    evaluator LLM that reviewed the predictions
    and identified areas for improvement.

Your task is to create an updated prompt that
    incorporates the feedback you received to
    improve the forecasting accuracy of the
    pipeline in the next round.
You can make minor changes, completely rewrite the
    prompt, or anything in between. Feel free to
    be creative, as long as you are taking into
    account the feedback you received.

Return only the text of the new prompt without any
    commentary. Do not include special variables
    in brackets or metadata.
Avoid overfitting to specific questions in the
    feedback, but incorporate any specific
    suggestions you received.

Return only the updated prompt, do not include any
    commentary.
```

Listing A.6: Prompt Update Prompt

```
You are part of an LLM pipeline that forecasts
   outcomes of real-world events and iteratively
   updates the prompt for forecasting to improve
   over time.
Your role is to summarize the feedback in this
   round, and the changes made to the prompt
   because of the feedback.
This will be used to keep track of the changes
   made over time.
Include information on
(1) what the feedback was (what did the evaluator
   think the predictions did well or poorly, and
   why),
(2) what changes were made to the prompt and why.

Keep it short (2 concise sentences maximum).
```

Listing A.7: Reflection Summary Prompt

```
You are part of a LLM pipeline that forecasts
   outcomes of real-world events and iteratively
   updates the prompt for forecasting to improve
   over time.
You are given a question, and you are to summarize
   the resolution of the question (what actually
   happened).
Give an in-depth answer with context. If there is
   not enough information, then say so.
```

Listing A.8: Resolution Summary Prompt

A.2 Representative Outputs

```
Question: Will Google Maps display Gulf of America
   instead of Gulf of Mexico before April 1, 2025?

Resolution Summary: "As of February 2025, Google
   Maps began displaying the Gulf of Mexico as
   the \"Gulf of America\" for users in the
   United States, following an executive order
   signed by President Donald Trump. This change
   was made official by updating the Geographic
   Names Information System (GNIS) in the U.S. to
   reflect the new name. Users in Mexico continue
   to see the label as \"Gulf of Mexico,\" while
   users in other countries see both names:
   \"Gulf of Mexico (Gulf of America)\"[1][3].

Given this information, it is clear that Google
   Maps did indeed start displaying \"Gulf of
   America\" instead of \"Gulf of Mexico\" for
   U.S. users before April 1, 2025. The rollout
   of this change began in early February 2025,
   aligning with the updates to official
   geographic naming systems in the U.S.[1][3].

However, it's worth noting that not all platforms
   or maps may adopt this change. For instance,
   the Newsweek article mentions that not all
   services, including possibly Apple Maps, may
   follow this renaming. Nonetheless, as of the
   current information, Google Maps has
   implemented the change for U.S. users.",

Citations: Bay News 9, Google Support, Google Blog
```

Listing A.9: Resolution Summary Example

```
Question: Will Google Maps display Gulf of America
   instead of Gulf of Mexico before April 1, 2025?
{
  "explanation": "To determine if Google Maps will
     change the name from Gulf of Mexico to Gulf
     of America, we need to consider several
     factors. First, we should check if there
     have been any official announcements or
     updates from Google regarding changes to
     their map services. Second, we should look
     for any political or social movements that
     might influence such a change, as the name
     change could be politically motivated.
     Lastly, we should investigate if there have
     been any recent discussions or controversies
     surrounding the naming of the Gulf of Mexico
     that could lead to a change.",
  "search_queries": [
      "Google Maps update",
      "Gulf of America news",
      "Gulf of Mexico controversy"
  ]
}
```

Listing A.10: Search Query Generation Example

```
<thinking>
The original question asks whether Google Maps will display
    "Gulf of America" instead of "Gulf of Mexico" before April
    1, 2025. To rephrase and expand this question, I would
    ask: "Considering the recent political directives and
    public discussions surrounding the renaming of
    geographical features, will Google Maps officially change
    the name of the Gulf of Mexico to Gulf of America prior to
    the date of April 1, 2025? What factors might influence
    this decision, and what evidence supports either outcome?"
Reasons why the answer might be no include:

Google Maps relies on authoritative data sources and official
    naming conventions, which currently recognize the Gulf as
    the Gulf of Mexico. (Strength: Strong)
There has been no official confirmation from Google regarding
    a timeline for implementing the name change, and recent
    articles indicate that the name remains unchanged as of
    now. (Strength: Strong)
The opposition to the name change from various stakeholders,
    including international entities, may deter Google from
    making the change. (Strength: Moderate)

Reasons why the answer might be yes include:

There is a political directive from former President Trump
    that suggests a name change, which could pressure Google
    to comply with governmental updates. (Strength: Moderate)
Public sentiment or political pressure could lead Google to
    adopt the name change to align with nationalistic
    sentiments. (Strength: Moderate)
If the U.S. government officially updates the name, Google may
    have no choice but to follow suit to maintain consistency
    in its mapping services. (Strength: Strong)

Aggregating these considerations, it appears that while there
    is political pressure and potential for change, the lack
    of official confirmation and reliance on authoritative
    naming conventions suggest that Google Maps may not
    implement the name change before the specified date. The
    strength of the reasons against the change outweighs those
    in favor, leading to a cautious outlook.
Based on the analysis, I would assign an initial probability
    of 0.30 to the likelihood that Google Maps will display
    "Gulf of America" instead of "Gulf of Mexico" before April
    1, 2025.
This calculated probability seems reasonably cautious, given
    the current lack of official confirmation and the reliance
    on established naming conventions. However, I should
    consider the potential for political influence and public
    sentiment, which could shift the landscape. Overall, I
    believe the initial probability is not excessively
    confident but could be slightly adjusted to account for
    the political climate.
My final prediction, reflecting a balanced consideration of
    the factors involved, is 0.35.
</thinking>
<probability>0.35</probability>",
```

Listing A.11: Forecast Generation Example (GPT 4o-mini, Initial Prompt)

```
"prediction_rationale": "<thinking>
1. **Rephrased and expanded question**:
Will Google Maps officially rename the \"Gulf of Mexico\" to
    \"Gulf of America\" in its mapping services prior to April
    1, 2025, following reports of political pressure from
    Donald Trump's administration, or will it maintain the
    current name based on existing geographical naming
    conventions and government databases?

2. **Reasons the answer might be \"No\"**:
   - **Official naming protocols**: Google Maps relies on
       authoritative sources like the U.S. Board on Geographic
       Names (BGN). If the BGN has not formally approved the
       change, Google is unlikely to act unilaterally.
       (Strength: 4/5)
   - **Conflicting media reports**: Credible outlets like the
       NYT and Al Jazeera explicitly state Google has no plans
       to change the name, while pro-Trump outlets (Fox, CNBC)
       report impending changes without evidence of official
       action. (Strength: 3/5)
   - **Historical precedent**: Google only updated \"Mount
       McKinley\" to \"Denali\" in 2015 after the BGN
       finalized the change. Similar bureaucratic inertia
       would likely apply here. (Strength: 4/5)

3. **Reasons the answer might be \"Yes\"**:
   - **Government directive compliance**: If Trump's order is
       formalized through executive action or BGN approval
       before April 2025, Google would likely comply quickly,
       as seen in past updates. (Strength: 2/5     no evidence
       of BGN movement)
   - **Political symbolism**: The Trump administration might
       prioritize this as a nationalist gesture, accelerating
       bureaucratic processes. (Strength: 1/5     low
       likelihood given procedural hurdles)

4. **Aggregated considerations**:
   - The U.S. government has not officially renamed the Gulf
       of Mexico as of the latest articles (Jan 2025).
   - Google's policy ties map labels to recognized
       authorities, not political statements.
   - Renaming a major geographical feature requires
       international recognition (e.g., UN, Mexico's
       government), which is absent here.
   - The April 2025 deadline allows only ~3 months for
       bureaucratic action, which is insufficient for such a
       contentious change.

5. **Initial probability**: 0.15 (15%)

6. **Calibration check**:
   - Base rate: Renaming major geographical features without
       international consensus is extremely rare.
   - Overconfidence risk: The 15% might underestimate the
       possibility of a rushed executive order. However,
       Mexico's likely diplomatic opposition and media
       skepticism (BBC, NYT) justify a low probability.
   - Additional factor: Apple Maps' parallel inaction (per
       Newsweek) reinforces institutional resistance to
       unilateral changes.

</thinking>
<probability>0.20</probability>"
```

Listing A.12: Forecast Generation Example (Deepseek R1, Initial Prompt)

```
<thinking> In considering whether Google Maps will display the
    name \"Gulf of America\" instead of \"Gulf of Mexico\"
    before April 1, 2025, it is essential to analyze the
    broader implications of this potential change. The
    renaming of geographical features can have significant
    societal, economic, and environmental impacts. For
    instance, a name change could influence national identity
    and cultural perceptions, as well as affect tourism and
    local economies that rely on the established name.
Historically, name changes of geographical features have often
    been contentious, reflecting deeper political and cultural
    conflicts. The proposal to rename the Gulf of Mexico has
    been associated with political figures and movements that
    seek to assert a particular narrative about American
    identity. This context suggests that public sentiment and
    political dynamics will play a crucial role in the outcome
    of this situation.

On the positive side, if Google Maps does implement the name
    change, it could be seen as a validation of certain
    political ideologies and a shift towards a more
    nationalistic narrative. This could resonate positively
    with specific demographics, potentially increasing user
    engagement with the platform. However, the negative
    outcomes could include backlash from those who oppose the
    name change, leading to public protests or calls for
    boycotts of Google services.

The uncertainty surrounding this issue is significant. Google
    has not provided a definitive timeline for the
    implementation of the name change, and there are no
    guarantees that the U.S. government will officially
    endorse the new name before the specified date.
    Additionally, public sentiment can shift rapidly,
    influenced by media narratives and political developments.

Given the historical context of similar events, such as the
    renaming of Mount McKinley to Denali, there is a precedent
    for such changes occurring, but they often take time and
    require broad consensus. Current trends indicate a growing
    polarization in public opinion regarding national
    identity, which could complicate the acceptance of a name
    change.

Taking all these factors into account, I would assign a
    probability of 0.65 to the likelihood that Google Maps
    will display "Gulf of America" instead of "Gulf of Mexico"
    before April 1, 2025. This reflects a moderate level of
    confidence, acknowledging both the potential for change
    and the significant uncertainties involved.
</thinking>

<probability>0.65</probability>",
```

Listing A.13: Forecast Generation Example (GPT 4o-mini, Final Prompt)

```
"prediction_rationale": "<thinking>

Source Analysis: Conflicting reports between outlets claiming
    imminent change (Fox Business, CNBC) vs. credible denials
    (NYT, Al Jazeera). T1 sources like NYT (Jan 29) explicitly
    state Google has no plans to change.
Political Context: Trump's order (if issued post-2024
    election) would fall under Transition Period Adjustments
    (0.7   ceiling). However, implementation requires formal
    USGS/BGN approval absent in current reporting.
Temporal Factors: Latest credible updates (Jan 30) confirm
    status quo. Governmental name change processes typically
    take 6-18 months, making April 2025 deadline unlikely
    without existing administrative action.
Corporate Policy: Google's historical pattern shows 3-5 month
    lag between official name changes and map updates. No
    evidence of formal BGN filing reduces probability.
Protocol Application: 45% confidence reduction applies due to
    unresolved political risk (unclear legal authority for
    unilateral renaming). Source cohesion penalty (-0.4  )
    from conflicting T2 reports.
</thinking>


<probability>0.17</probability>",
```

Listing A.14: Forecast Generation Example (Deepseek R1, Final Prompt)

```
In reflecting on my past predictions, several key observations emerge that can
     inform improvements in my forecasting approach.

1. **Recurring Patterns in Reasoning**: A notable pattern in my reasoning was a
     tendency to rely heavily on historical behavior and established patterns of
     political discourse, particularly in the context of Donald Trump's speeches.
     While this approach is grounded in a solid understanding of past events, it
     sometimes led to overconfidence in predictions. For instance, I consistently
     assigned high probabilities (around 0.85) to the likelihood of Trump
     mentioning specific figures like Biden, AOC, or Pelosi based on their
     historical significance in his rhetoric. This reliance on historical patterns
     may have overshadowed the need to consider the evolving political landscape
     and public sentiment, which can shift rapidly.

2. **Evidence and Its Impact**: The articles I used provided valuable insights into
     Trump's past behavior and the political dynamics at play. However, I noticed
     that I could have better integrated qualitative assessments of public
     sentiment and current events. For example, while I referenced Trump's
     historical criticisms of Biden, I did not sufficiently account for potential
     shifts in public opinion or the political climate leading up to the address.
     This oversight may have contributed to a lack of nuance in my predictions,
     particularly in cases where public sentiment was trending towards
     bipartisanship or unity.

3. **Superforecasting Techniques**: I employed some superforecasting techniques,
     such as considering both positive and negative outcomes and addressing
     counterarguments. However, I could improve by explicitly discussing the
     implications of public sentiment and political dynamics on the likelihood of
     specific mentions in speeches. For instance, in the case of Trump's address, I
     could have explored how recent events or media narratives might influence his
     choice of rhetoric, which would have added depth to my analysis.

4. **Calibration of Confidence Levels**: My confidence levels tended to be
     systematically high, particularly in politically charged scenarios. This
     overconfidence was evident in my consistent assignment of probabilities around
     0.85 for mentions of key political figures. While I justified these
     probabilities based on historical patterns, I did not sufficiently account for
     the inherent unpredictability of political contexts. Moving forward, I should
     aim for a more calibrated approach, perhaps by incorporating a wider range of
     probabilities that reflect the uncertainties involved.

5. **Diverse Categories of Questions**: I noticed that my reasoning varied across
     different categories of questions, particularly when addressing political
     figures versus broader policy issues. In cases involving specific individuals,
     I often defaulted to historical behavior without adequately considering the
     broader implications of their current political standing or public perception.
     This inconsistency suggests a need for a more systematic approach to reasoning
     across different domains, ensuring that I consistently evaluate the relevance
     of historical context, current trends, and potential uncertainties.

6. **Integration of Public Sentiment**: One area for improvement is the explicit
     integration of public sentiment and political dynamics into my forecasts.
     While I acknowledged these factors, I did not always translate them into my
     probability assessments effectively. For example, understanding how public
     sentiment towards Biden or Pelosi might shift in response to current events
     could have influenced my predictions regarding Trump's rhetoric. A more
     structured approach to evaluating public sentiment and its potential impact on
     political discourse would enhance the robustness of my forecasts.

7. **Addressing Counterarguments**: While I made efforts to consider
     counterarguments, there were instances where I could have delved deeper into
     alternative perspectives. For example, in predicting whether Trump would
     mention Biden, I could have explored the possibility that he might choose to
     focus on broader themes to appeal to a wider audience, rather than engaging in
     direct criticism. A more thorough examination of potential counterarguments
     would provide a more balanced view and improve the accuracy of my predictions.

In summary, my reflections highlight the importance of balancing historical
     patterns with an awareness of current political dynamics and public sentiment.
     Moving forward, I aim to enhance my forecasting accuracy by integrating a
     wider range of evidence, calibrating my confidence levels more effectively,
     and systematically addressing uncertainties and counterarguments. By doing so,
     I hope to improve the overall robustness and reliability of my predictions
     across various categories of questions.
```

Listing A.15: GPT 4o-mini, Reflection Example

```
Reflection on Forecasting Performance

Overconfidence in Historical Patterns:
Both predictions exhibited systematic overconfidence (75% and 68% probabilities)
     despite lacking direct evidence of the mentions occurring. This stemmed from
     overweighting Trump's historical tendency to criticize political opponents
     (1.25   domain history multiplier) while underweighting:

Post-election context shifts (Nov 2024 Trump-Biden meeting showing cordiality)
Formal congressional address norms (agenda-setting vs. campaign rally rhetoric)
Strategic silence incentives for symbolic first addresses

The protocol's temporal decay (40% reduction for >30-day-old geopolitical
     references) proved insufficient to counteract anchoring bias from 2024
     behavioral patterns. This highlights a need for event-type-specific decay
     rates, with accelerated penalties (e.g., 3.5% daily) for political criticism
     predictions post-power transitions.
Contra-Indicator Miscalibration:
Recent contra-indicators (2024 Trump-Biden transition meeting, Pelosi/AOC committee
     tensions) received only 15% reductions despite directly signaling reduced
     adversarial urgency. The protocol needs:

Dynamic contra-indicator weighting (up to 35% penalty for direct behavioral shifts)
Interaction terms between recent diplomatic events and criticism likelihood
Explicit "strategic silence" clauses for inaugural policy speeches

Strategic Context Mis:
The 1.15   symbolic action multiplier was misapplied by conflating agenda-setting
     symbolism with adversarial rhetoric. Required refinements:

Subtype differentiation in strategic context (0.6   penalty for opponent mentions
     in policy-focused addresses)
Forward-looking agenda filters reducing criticism weights when speech outlines new
     initiatives

Source Tier Limitations:
Reliance on T2/T3 sources (media analyses vs. direct transcripts) without recent T1
     validation (post-2024 election dynamics) went unflagged. Improvements needed:

Recency-tiered credibility: 0.5   penalty for political behavior predictions
     lacking T1 sources <90 days pre-event
Cohesion scoring for multi-source clusters about behavioral shifts

Confidence Band Errors:
The 35% confidence reduction for unresolved risks was inadequate against structural
     political. Calibration requires:

Asymmetric uncertainty bands: +25%/-40% for political criticism predictions
Election cycle adjustments: 0.7   ceiling for post-transition period rhetoric
     forecasts

These patterns suggest systemic issues in handling regime change dynamics and
     formal speech conventions. Protocol v7.2's focus on technical/historical
     multipliers without political phase detection created blind spots. Future
     iterations need administrative context layers that adjust weights based on
     officeholder status (incumbent vs. returning) and speech type (inaugural vs.
     mid-term).
```

Listing A.16: Deepseek R1, Reflection Example

```
"Instructions:

1. When forecasting a real-world event, rephrase and expand the question to ensure
     clarity and comprehensiveness, while maintaining all relevant information from
     the original question. Consider the broader implications and context of the
     event, including potential societal, economic, and environmental impacts.

2. Analyze potential reasons for both negative and positive outcomes, providing
     well-reasoned arguments for each. Rate the strength of each reason based on
     your knowledge and the context provided, and be explicit about uncertainties
     involved. Address counterarguments thoroughly and consider alternative
     perspectives, ensuring a balanced view of optimism and caution.

3. Incorporate historical context and base rates of similar events, while also
     integrating current trends and external factors that may influence the
     outcome. Think critically and holistically about the factors at play, ensuring
     a comprehensive analysis.

4. Output an initial probability (prediction) based on the analysis conducted in
     steps 1-3, ensuring it is a single value between 0 and 1. Reflect on the
     calibration of your confidence level, considering whether it is excessively
     confident or not confident enough, and strive for consistency in assessing
     uncertainty across different domains.

5. Evaluate the confidence level of your calculated probability, considering any
     additional factors that may influence the forecast. Ensure that your reasoning
     is thorough and well-structured, explicitly addressing uncertainties and
     counterarguments.

6. Output your final prediction as a single probability value, formatted with an
     asterisk at the beginning and end of the decimal. Avoid providing a range of
     probabilities; always return a single probability value.

7. Ensure a systematic approach to reasoning across different domains, integrating
     external influences, historical context, and current trends consistently.
     Address the inherent unpredictability in political contexts and adjust
     confidence levels accordingly.

8. Reflect on past predictions to identify patterns and insights that can inform
     improvements in forecasting accuracy, focusing on calibration of confidence
     levels, evidence integration, and thorough evaluation of counterarguments.
     Emphasize the importance of balancing optimism with caution and critically
     assessing the evidence relied upon.

9. Strive for consistency in your approach across various categories of questions,
     ensuring that you systematically consider historical context, current trends,
     and potential uncertainties in all forecasts. Aim to integrate a wider range
     of evidence, including expert opinions and emerging trends, to create a more
     comprehensive picture.

10. Pay special attention to the integration of quantitative data where applicable,
     ensuring that your analysis is supported by specific figures and statistics to
     enhance the robustness of your predictions. Additionally, ensure that you
     explicitly address the implications of public sentiment and political dynamics
     in your analysis.

11. Incorporate a section that explicitly discusses the potential impact of public
     sentiment and political dynamics on the event being forecasted, ensuring that
     these factors are integrated into the overall analysis and probability
     assessment.

12. Be explicit about the uncertainties and potential shifts in strategy or
     external circumstances that could disrupt established patterns, particularly
     in cases where historical performance is relied upon.

13. Enhance the analysis of public sentiment and political dynamics by discussing
     how these factors might influence consumer behavior and corporate strategies,
     particularly in sensitive contexts.

14. Ensure that your reasoning is balanced, integrating both the potential for
     success and the risks involved in each prediction.

15. Emphasize the need for a thorough examination of how current events and media
     narratives influence public opinion, and how these shifts may impact the
     outcomes being forecasted.

16. Encourage a more structured approach to explicitly address counterarguments and
     uncertainties in the analysis, ensuring a comprehensive evaluation of all
     relevant factors.

Return your answer in XML format. Use <thinking> tags to insert your thoughts.
Use <probability> tag to insert your final prediction, and nothing else.
No bolding or other formatting. For instance, <thinking>...</thinking>
     <probability>0.95</probability>
This is crucial: Never return a range of probabilities, such as "Between 0.90 and
     0.95". Always return a single probability, such as "0.90"."
```

Listing A.17: GPT 4o-mini, Final Prompt Example

```
Updated Forecasting Protocol v7.3

Political Phase Detection

Transition Period Adjustments:
0.7   ceiling for predictions during power transitions (90 days
     pre/post-inauguration)
3.5% daily decay rate for political criticism predictions post-election
Officeholder Status:
1.15   weight for incumbent rhetoric patterns
0.6   penalty for adversarial predictions against former opponents during policy
     addresses


Strategic Context Subtypes

Symbolic Action Differentiation:
1.3   for legacy-building initiatives
0.55   for opponent criticism in legislative-focused speeches
Agenda Filters:
40% reduction in adversarial weights when speech outlines >2 new policy initiatives


Dynamic Contra-Indicator System

Behavioral Shift Multipliers:
Up to 35% penalty for recent diplomatic breakthroughs (<45 days)
1.4   interaction boost when technical indicators align with leadership changes
Strategic Silence Clause:
0.25   criticism probability for inaugural/state-of-union addresses


Temporal-Source Validation

Recency-Tiered Credibility:
0.5   penalty for political predictions lacking T1 sources <90 days pre-event
T1 Decay: 12% monthly (20% during election cycles)
Cohesion Scoring:
1.25   boost when >3 T2+ sources confirm behavioral shifts


Asymmetric Confidence Framework

Political Uncertainty Bands:
+22%/-38% for criticism predictions vs  15 % technical events
Speech-Type Adjustments:
0.8   maximum confidence for post-transition ceremonial addresses


Execution Calculus
[((Phase_Adjustment    Agenda_Filter)^(1 + 0.15  subtype_weight  ))
   (1 - min(Strategic_Silence, 0.75))
   (Source_Cohesion^(2 - 0.4  contra_penalties  ))
   (Temporal_Decay + 0.07  officeholder_status  )
   (Asymmetric_Band    1.18)]
/ (Initiative_Complexity    1.4 + 0.05  transition_decay  )

Apply 45% confidence reduction for predictions with unresolved political risk
     flags. All speech-type predictions require T1 transcript analysis <72hr
     pre-event. Ceremonial address clauses override technical indicators when
     protocol age <120 days. Final scores normalized with Tukey biweight against
     election-cycle benchmarks.
Return your answer in XML format. Use <thinking> tags to insert your thoughts.
Use <probability> tag to insert your final prediction, and nothing else.
No bolding or other formatting. For instance, <thinking>...</thinking>
     <probability>0.95</probability>
This is crucial: Never return a range of probabilities, such as "Between 0.90 and
     0.95". Always return a single probability, such as "0.90".
```

Listing A.18: Deepseek R1, Final Prompt Example

```
The feedback highlighted a tendency towards
   overconfidence in predictions due to reliance
   on historical patterns, and insufficient
   integration of public sentiment and current
   events. Consequently, the prompt was updated
   to emphasize the need for a balanced approach,
   incorporating a wider range of evidence,
   including public sentiment and political
   dynamics, to enhance prediction accuracy.
```

Listing A.19: GPT 4o Mini, Reflection Summary Example

```
(1) The feedback highlighted overconfidence in
   historical patterns and inadequate handling of
   political transitions, leading to
   miscalibrated predictions. (2) Changes to the
   prompt include implementing political phase
   detection, strategic context subtypes, and a
   dynamic contra-indicator system to better
   account for political transitions and speech
   types.
```

Listing A.20: Deepseek R1, Reflection Summary Example

Bibliography

Anantaprayoon, P., Kaneko, M., & Okazaki, N. (2025, March). Intent-Aware Self-Correction for Mitigating Social Biases in Large Language Models [arXiv:2503.06011 [cs]]. https://doi.org/10.48550/arXiv.2503.06011

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*(4), 559–573. https://doi.org/10.1016/0169-2070(92)90066-I

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023, August). ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate [arXiv:2308.07201 [cs]]. https://doi.org/10.48550/arXiv.2308.07201

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526. https://doi.org/10.1017/S1930297500004599

Chen, X., Lin, M., Schärli, N., & Zhou, D. (2023, October). Teaching Large Language Models to Self-Debug [arXiv:2304.05128 [cs]]. https://doi.org/10.48550/arXiv.2304.05128

Chia, Y. K., Chen, G., Tuan, L. A., Poria, S., & Bing, L. (2023, November). Contrastive Chain-of-Thought Prompting [arXiv:2311.09277 [cs]]. https://doi.org/10.48550/arXiv.2311.09277

De Baets, S., & Vanderheyden, K. (2021). Individual differences in the susceptibility to forecasting biases [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.38 *Applied Cognitive Psychology*, *35*(4), 1106–1114. https://doi.org/10.1002/acp.3831

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025, January). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [arXiv:2501.12948 [cs]]. https://doi.org/10.48550/arXiv.2501.12948

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023, May). Improving Factuality and Reasoning in Language Models through Multiagent Debate [arXiv:2305.14325 [cs]]. https://doi.org/10.48550/arXiv.2305.14325

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance*, *29*(3), 352–369. https://doi.org/10.1016/0030-5073(82)90250-1

Halawi, D., Zhang, F., Yueh-Han, C., & Steinhardt, J. (2024, February). Approaching Human-Level Forecasting with Language Models [arXiv:2402.18563 [cs]]. https://doi.org/10.48550/arXiv.2402.18563

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024, March). Large Language Models Cannot Self-Correct Reason-

ing Yet [arXiv:2310.01798 [cs]]. https://doi.org/10.48550/arXiv.2310.01798

Hwang, H., Kim, D., Kim, S., Ye, S., & Seo, M. (2024, November). Self-Explore: Enhancing Mathematical Reasoning in Language Models with Fine-grained Rewards. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 1444–1466). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.78

Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., & Ren, X. (2021, August). ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4636–4650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.357

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024, December). When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs [arXiv:2406.01297 [cs]]. https://doi.org/10.1162/tacl_a_00713/125177

Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2024, April). A Survey of Reinforcement Learning from Human Feedback [arXiv:2312.14925 [cs]]. https://doi.org/10.48550/arXiv.2312.14925

Lee, K.-H., Fischer, I., Wu, Y.-H., Marwood, D., Baluja, S., Schuurmans, D., & Chen, X. (2025, January). Evolving Deeper LLM Thinking [arXiv:2501.09891 [cs]]. https://doi.org/10.48550/arXiv.2501.09891

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5

Liu, G., Qi, Z., Zhang, X., Cheng, L., & Johnson, K. M. (2025, March). Self-correction is Not An Innate Capability in Large Language Models: A Case Study of Moral Self-correction [arXiv:2410.20513 [cs]]. https://doi.org/10.48550/arXiv.2410.20513

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (n.d.). SELF-REFINE: Iterative Refinement with Self-Feedback.

OpenAI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., … Li, Z. (2024, December). OpenAI o1 System Card [arXiv:2412.16720 [cs]]. https://doi.org/10.48550/arXiv.2412.16720

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, March). Training language models to follow instructions with human feedback [arXiv:2203.02155 [cs]]. https://doi.org/10.48550/arXiv.2203.02155

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., & Tramèr, F. (2025, January). Consistency Checks for Language Model Forecasters [arXiv:2412.18544 [cs]]. https://doi.org/10.48550/arXiv.2412.18544

Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies [Place: Cambridge, MA Publisher: MIT Press]. *Transactions of the Association for Computational Linguistics*, *12*, 484–506. https://doi.org/10.1162/tacl_a_00660

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023, March). Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback [arXiv:2302.12813 [cs]]. https://doi.org/10.48550/arXiv.2302.12813

Pratt, S., Blumberg, S., Carolino, P. K., & Morris, M. R. (2024, June). Can Language Models Use Forecasting Strategies? [arXiv:2406.04446 [cs]]. https://doi.org/10.48550/arXiv.2406.04446

Qu, Y., Zhang, T., Garg, N., & Kumar, A. (2024, July). Recursive Introspection: Teaching Language Model Agents How to Self-Improve [arXiv:2407.18219 [cs]]. https://doi.org/10.48550/arXiv.2407.18219

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024, July). Direct Preference Optimization: Your Language Model is Secretly a Reward Model [arXiv:2305.18290 [cs]]. https://doi.org/10.48550/arXiv.2305.18290

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., & Fawzi, A. (2024). Mathematical discoveries from program search with large language models [Publisher: Nature Publishing Group]. *Nature*, *625*(7995), 468–475. https://doi.org/10.1038/s41586-023-06924-6

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025, March). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [arXiv:2402.07927 [cs]]. https://doi.org/10.48550/arXiv.2402.07927

Schoenegger, P., & Park, P. S. (2023, October). Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament [arXiv:2310.13014 [cs]]. https://doi.org/10.48550/arXiv.2310.13014

Schoenegger, P., Park, P. S., Karger, E., Trott, S., & Tetlock, P. E. (2025). AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. *ACM Transactions on Interactive Intelligent Systems*, *15*(1), 1–25. https://doi.org/10.1145/3707649

Schoenegger, P., Tuminauskaite, I., Park, P. S., & Tetlock, P. E. (2024, July). Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy [arXiv:2402.19379 [cs]]. https://doi.org/10.48550/arXiv.2402.19379

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037/0096-3445.130.3.380

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023, October). Reflexion: Language Agents with Verbal Reinforcement Learning [arXiv:2303.11366 [cs]]. https://doi.org/10.48550/arXiv.2303.11366

Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D., & Pretorius, A. (2024, July). Should we be going MAD? A Look at Multi-Agent Debate

Strategies for LLMs [arXiv:2311.17371 [cs]]. https://doi.org/10.48550/arXiv.2311.17371

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction* [Pages: 340]. Crown Publishers/Random House.

Turtel, B., Franklin, D., & Schoenegger, P. (2025, February). LLMs Can Teach Themselves to Better Predict the Future [arXiv:2502.05253 [cs]]. https://doi.org/10.48550/arXiv.2502.05253

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023, March). Self-Consistency Improves Chain of Thought Reasoning in Language Models [arXiv:2203.11171 [cs]]. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [arXiv:2201.11903 [cs]]. https://doi.org/10.48550/arXiv.2201.11903

Xiong, W., Zhang, H., Ye, C., Chen, L., Jiang, N., & Zhang, T. (2025, February). Self-rewarding correction for mathematical reasoning [arXiv:2502.19613 [cs]]. https://doi.org/10.48550/arXiv.2502.19613

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023, March). ReAct: Synergizing Reasoning and Acting in Language Models [arXiv:2210.03629 [cs]]. https://doi.org/10.48550/arXiv.2210.03629

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., & Wang, W. (2024, July). MIRAI: Evaluating LLM Agents for Event Forecasting [arXiv:2407.01231 [cs]]. https://doi.org/10.48550/arXiv.2407.01231

Yu, Z., He, L., Wu, Z., Dai, X., & Chen, J. (2023, October). Towards Better Chain-of-Thought Prompting Strategies: A Survey [arXiv:2310.04959 [cs]]. https://doi.org/10.48550/arXiv.2310.04959

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., & Hendrycks, D. (2022, October). Forecasting Future World Events with Neural Networks [arXiv:2206.15474 [cs]]. https://doi.org/10.48550/arXiv.2206.15474