

AIA Forecaster: Technical Report

Rohan Alur*, Bradly C. Stadie*, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, Jasjeet S. Sekhon

*Bridgewater AIA Labs
New York, NY*

Abstract

This technical report describes the AIA FORECASTER, a Large Language Model (LLM)-based system for judgmental forecasting using unstructured data. The AIA FORECASTER approach combines three core elements: agentic search over high-quality news sources, a supervisor agent that reconciles disparate forecasts for the same event, and a set of statistical calibration techniques to counter behavioral biases in large language models. On the FORECASTBENCH benchmark (Karger et al., 2024), the AIA FORECASTER achieves performance equal to human superforecasters, surpassing prior LLM baselines. In addition to reporting on FORECASTBENCH, we also introduce a more challenging forecasting benchmark sourced from liquid prediction markets. While the AIA FORECASTER underperforms market consensus on this benchmark, an ensemble combining AIA FORECASTER with market consensus outperforms consensus alone, demonstrating that our forecaster provides additive information. Our work establishes a new state of the art in AI forecasting and provides practical, transferable recommendations for future research. To the best of our knowledge, this is the first work that verifiably achieves expert-level forecasting at scale.

* These authors contributed equally to this work. Correspondence to aialabs@bwatert.com.

Contents

1	Introduction	3
2	Related Work	5
3	Methodology	7
3.1	Notation and Preliminaries	7
3.2	The AIA FORECASTER	8
4	The AIA Forecaster Is Indistinguishable from Superforecasters	10
4.1	Benchmarks and Evaluation Metrics	10
4.2	Forecaster Performance	12
5	Understanding Search in Forecasting	13
5.1	Search Is Critical for Forecasting	14
5.2	Mitigating Foreknowledge Bias in Search	17
5.3	LLM Forecasts Are Diversifying to Market Prices	18
6	Search Is Critical for Live Markets	19
7	Ensembling, Synthesis, and Statistical Corrections	21
7.1	Simple Averaging Is a Strong Baseline	21
7.2	Supervisor Agents Resolve Disagreements	23
7.3	Statistical Corrections and Where They Help	24
8	Discussion	27
Appendix		34
A.	Uncertainty Quantification for Primary Results	34
B.	Evaluating Different Base Models	35
C.	Sources of Foreknowledge Bias	36
D.	Foreknowledge Analysis Prompt	37
E.	Further Search Details and Results	39
F.	MARKETLIQUID Dataset Construction	40
G.	LLMs Attenuate Forecasts	41
G.1	Examples of Hedging Behavior	41
G.2	Platt Scaling is Generalized Log Odds Extremization	42

1 Introduction

Forecasting is a universal problem. Any system that makes informed judgments about what may happen in the future relies on at least some degree of forecasting. In agriculture, this takes the form of predicting future crop yields and food system resilience (Bueechi et al., 2023; Tanaka et al., 2023; Paudel et al., 2021). In biotechnology and medicine, clinical trials are carefully designed to help researchers predict the impact of newly developed treatments (Qian et al., 2025; Curth et al., 2024). Elsewhere in biology, AlphaFold can be viewed as a form of forecasting protein folding structure under uncertainty (Jumper et al., 2021). In academia, grant funding is a form of forecasting which scientific discoveries will be made conditional on funding (Tohalino and Amancio, 2022). Climate and environmental science models local and long-term weather trends, which are of universal importance (Soliman, 2024; Price et al., 2025; Google DeepMind, 2024). In political science, forecasting election outcomes and the impact of policy decisions are an essential component of political calculus (Jennings et al., 2020; Wang et al., 2015; Gelman et al., 2020). Militaries make use of forecasting for threat assessment and conflict predictions (Zequeira, 2024; Probasco et al., 2025). Forecasting is the very heartbeat of financial firms, which use it to estimate future economic conditions in an uncertain world (Tsay, 2005). Wherever there is uncertainty about the future, there is forecasting.

The literature distinguishes between two high-level approaches to forecasting. The first, statistical forecasting, utilizes some combination of tabular data and simulation to develop a mathematical model of what might occur in the future. In contrast, *judgmental forecasting* is the process of aggregating unstructured data (e.g., news articles, scientific reports, etc.) and using this data in conjunction with past experience to logically predict some future outcome (Lawrence et al., 2006).

To better understand the two types of forecasting and their relationship, let us consider an example. Suppose a business wants to predict whether a certain piece of real estate will appreciate in value next year. By comparing the property to other similar properties, and looking at recent trends in the market, the company can develop a statistical model with some degree of sophistication. However, this modeling has its limits. Suppose that a news article is released, claiming that the foundation of a nearby building is shifting, causing it to slowly tilt. Even without performing additional research, a human can predict that this fact will likely put downward pressure on the property value. However, coercing this information into a form amenable to mathematical analysis is difficult. Thus, an ideal forecast would employ some abstract reasoning process which aggregates unstructured data, past experience, and logical thinking (i.e., judgment) to arrive at a sound conclusion.

This paper will focus exclusively on the judgmental forecasting problem (hereafter referred to simply as “forecasting”). Recently, there has been an uptick in research which studies the judgmental forecasting problem, due in large part to the ubiquity of Large Language Models (LLMs). By giving LLMs access to news databases and asking them to forecast future events, it is possible to systematically evaluate the forecasting abilities of machine

intelligence at scale. Several recent papers have endeavored to create and evaluate LLM-based forecasters (Halawi et al., 2024; Karger et al., 2024; Schoenegger et al., 2024; Turtel et al., 2025; Schoenegger et al., 2025). The rapid advances of LLMs on reasoning tasks (e.g., achieving gold medal performance at the International Math Olympiad (Google DeepMind, 2025)) offer reasons for optimism, and prompt the obvious question:

Can LLMs match or even exceed the performance of expert human forecasters?

Contributions. We answer this question in the affirmative by introducing the AIA FORECASTER. On FORECASTBENCH (Karger et al., 2024), the most widely used academic benchmark for forecasting, we achieve results that are statistically indistinguishable from human superforecasters. We also introduce a new benchmark which consists of 1610 questions sourced from a popular public prediction market platform. These questions cover decision-relevant events sourced from liquid markets, and are significantly harder than those on FORECASTBENCH. On this benchmark, the AIA FORECASTER lags slightly behind market consensus. However, an ensemble of the system’s forecasts with the market consensus outperforms consensus alone, indicating that the AIA FORECASTER is diversifying with respect to market prices. We corroborate these results by establishing a track record of real-time, forward-looking forecasts. In this modest sample, the AIA FORECASTER is highly competitive with market consensus in liquid prediction markets.

In addition to our core experiments, we also conduct extensive analysis to understand what makes a good forecaster. Our methodological contributions are as follows:

1. The AIA FORECASTER is equipped with the ability to perform a robust search over a large corpus of high-quality documents. Critically, this search is both *agentic* and *adaptive* in nature: the model is granted full discretion over what to search, and can adaptively seek out new information by conditioning on the results of earlier queries. We show there is a deep relationship between search quality and forecast quality, resolving a major open question in the literature.
2. We conduct an exhaustive analysis of *foreknowledge bias*, or unintended violations of the intended information cutoff. In particular, ex-post evaluation of such systems requires the ability to “travel back in time” and feed the forecaster news “as of” the intended forecasting date. Reliably enforcing this information cutoff is a surprisingly challenging problem (Paleka et al., 2025). We build a custom pipeline for detecting foreknowledge bias and demonstrate its efficacy, confirming the integrity of our results. We also introduce a live benchmark, MARKETNIGHTLY, which includes only active, liquid prediction markets, and thus precludes the use of foreknowledge. Results on this benchmark corroborate our other findings.
3. We show that individual forecast runs are unstable, and that ensembling over many forecasts is absolutely essential to obtain frontier-level forecasting performance. We

also show that the exact method of ensembling is crucial. It is not enough to take the simple mean of forecasts. Instead, the AIA FORECASTER passes the results of individual forecasting agents to a supervisor agent, which performs additional search to correct mistakes and reconcile differing perspectives.

4. LLMs are fundamentally miscalibrated for probabilistic prediction under uncertainty. They hedge too much and are too cautious. We examine a variety of different statistical correction techniques that can be applied to correct these biases. We derive a mathematical connection, which has not yet been documented in the literature, between Platt scaling and the extremization of forecasts. We also show that simple extremization techniques are often sufficient to obtain large gains in forecasting performance.

Organization. In Section 2, we provide an overview of related work. In Section 3, we describe the architecture of the AIA FORECASTER, and in Section 4 we present our main results on multiple variants of FORECASTBENCH and our novel MARKETLIQUID benchmark. In Section 5 we investigate the role of search, and address methodological challenges related to *foreknowledge bias*. We complement these results in Section 6, where we study real-time forecasting performance in live prediction markets. In Section 7, we study a range of ensembling and statistical calibration techniques, which we show can substantially improve on the performance of a single forecasting agent. Finally, in Section 8, we conclude with a discussion of limitations, open questions, and future work.

2 Related Work

We focus on *judgmental forecasting*, in which a forecaster is tasked with producing the probability that a binary event will occur (Lahiri and Yang, 2013; Tetlock and Gardner, 2016; Tetlock, 2017; Hanson, 1999). In the past 30 years, judgmental forecasting has become an increasingly important area of study (Lawrence et al., 2006). Two major findings that have increased its salience are (1) that the *wisdom of crowds* (aggregated predictions across many forecasters) often beats experts in judgmental forecasting (Surowiecki, 2004; Bassamboo et al., 2015; Da and Huang, 2019; Lichtendahl Jr et al., 2013), and (2) the existence of *superforecasters*, who consistently outperform even the wisdom of crowds (Tetlock, 2017; Tetlock and Gardner, 2016). The literature demonstrates that superforecasters are rare, and even subject-matter experts are often poor forecasters in their area of expertise. What’s more, superforecasting ability in one domain does not necessarily transfer to other domains.

With the advent of strong AI systems, there has been increased interest in AI forecasting (Halawi et al., 2024). AI forecasting is highly scalable, so if the performance of AI forecasters matches or exceeds superforecasters, these systems can be widely deployed. Of course, in adversarial settings (e.g., financial markets), widely available AI forecasters

may in turn increase the difficulty of subsequent prediction tasks, requiring systems which continuously evolve.¹

Techniques for AI forecasters. Modern AI forecasters provide context to an LLM and prompt it to produce the probability that a future event will occur (Wang et al., 2024; Halawi et al., 2024; Schoenegger et al., 2024). However, as with any modern AI system, there are many research choices that dramatically affect performance. For example, the LLM’s prompt (Schoenegger et al., 2025) and the news provided to the LLM (Wang et al., 2024) have both been shown to carry an outsize impact on AI’s forecasting ability (Turtel et al., 2025).

Benchmarks for AI forecasters. Recently, several benchmarks have been developed to evaluate the prediction quality of AI forecasters. Broadly, all benchmarks can be partitioned into two categories: *static* and *dynamic*. Static benchmarks are gathered at a single point in time and held fixed indefinitely thereafter. Dynamic benchmarks are continually updated to include fresh questions on a regular basis.

Examples of static benchmarks include Halawi et al. (2024), along with many others (Jin et al., 2020; Mutschlechner and Jatowt, 2025; Yuan et al., 2025; Wang et al., 2025; Ye et al., 2024; Zou et al., 2022). Static benchmarks will always atrophy as time marches on and new LLMs are trained. This is because newer models have knowledge of the world which is acquired after the question was initially posed. For example, if a benchmark published in 2023 asks the LLM to forecast who will win the 2024 US presidential election, models trained at later dates may acquire knowledge of the outcome (and other closely related events) through their pre-training corpus (Sarkar and Vafa, 2024). Models may also be given access to tools which inadvertently leak knowledge beyond the intended information cutoff (Paleka et al., 2025). The foreknowledge bias problem, together with several other issues related to the quality of baseline forecasts², is often the limiting factor in using static benchmarks to evaluate forecasting systems.

Regarding dynamic benchmarks, there are three major choices at present: FORECASTBENCH (Karger et al., 2024), Prophet Arena (Tao et al., 2025), and the Metaculus AI Forecasting Benchmarking Series (Metaculus, 2025). FORECASTBENCH sources questions from public prediction markets, including Manifold, Polymarket, and Metaculus. It filters out “low-quality” questions using an LLM. Given a list of questions, it then produces “combination questions”, which ask the model to forecast boolean conjunctions of the original events. Prophet Arena also sources questions from prediction markets, but fixes the news sources the LLMs can use. This simplifies the forecasting task significantly because, as we argue in Section 5, effective search is arguably the most critical component of effective forecasting. The Metaculus AI Forecasting Benchmarking Series releases

-
1. More generally, tasks in which predictions influence the distribution of future outcomes are known as *performative* prediction tasks (Perdomo et al., 2020).
 2. In particular, if one does not collect high-quality superforecaster or crowd predictions at the time the benchmark is created, it is impossible to travel back in time and obtain these baseline forecasts. Many extant benchmarks lack such baselines, which are crucial for evaluation.

questions on a weekly basis; this benchmark only offers a very small subset (48 questions) on which to evaluate LLMs against human performance.

A major issue with all three dynamic benchmarks is domain-specific relevancy. For example, FORECASTBENCH includes questions on the ELO scores of professional chess players, and the Metaculus AI Forecasting Benchmarking Series includes questions on podcast rankings. To the best of our knowledge, these questions are irrelevant from a policy or economic perspective.

3 Methodology

3.1 Notation and Preliminaries

In this work, we consider judgmental binary forecasting, wherein an LLM is presented with some binary question q . Toward answering the question q in an informed manner, we allow the LLM to gather some evidence, \mathcal{E} , before attempting its answer. For example, we might query an online news database of articles for articles related to q .

Using the gathered evidence, the LLM makes a prediction p , which denotes the probability of the outcome under scrutiny occurring. If the LLM is absolutely sure the event in question will come to pass, it should set $p = 1.0$. Likewise, if the LLM is sure the event will not happen, it should set $p = 0.0$. Intermediate values such as $p = 0.3$ should be used if the LLM is uncertain about the final outcome, but nonetheless has some evidence which shades its beliefs in a particular direction.

This process of mulling over evidence and assigning a probability can be denoted as

$$\pi : (q, \mathcal{E}) \rightarrow p$$

Of course, we have hidden a great many difficulties with the explanation above. Principal among them, we supposed the existence of an evidence gathering pipeline \mathcal{E} , tasked with collecting information that is useful for informing the LLM’s predictions. As good human researchers know, evidence gathering is often a highly iterative process: some article contains a key piece of information, which sparks further research, cascading in dozens of queries to find the information required to fully understand the underlying situation. Human forecasters may also work collaboratively, exchanging information and engaging in debate to arrive at a final forecast.

More generally, in designing an LLM forecaster, we face a vast array of design decisions, many of which mirror those faced by expert human forecasters. These choices include the choice of retrieval pipeline, the method of reasoning over evidence, and the manner in which to evaluate performance, among others. Below we describe the design of the AIA FORECASTER, which takes inspiration from the human forecasting process.

3.2 The AIA Forecaster

Architecture. The AIA FORECASTER is a multi-agent forecasting system, in which several agents perform an adaptive search procedure to retrieve relevant information, reason over the retrieved content, and finally synthesize this reasoning to produce a single probability. The system then reconciles these independent forecasts and applies a statistical correction to produce a final answer.

A full visualization of our forecasting pipeline can be found in Figure 1. In the pipeline, we see that M individual agents are spawned. Each agent carries out independent research and produces an initial forecast probability. Those individual forecasts are passed to a supervisor agent, which identifies disagreements among the M agents and is given a final chance to query our news database with any clarifying questions. After clarifying, the reconciliation pipeline outputs an aggregated forecast, which is then passed through a Platt scaling step to produce the final probability.

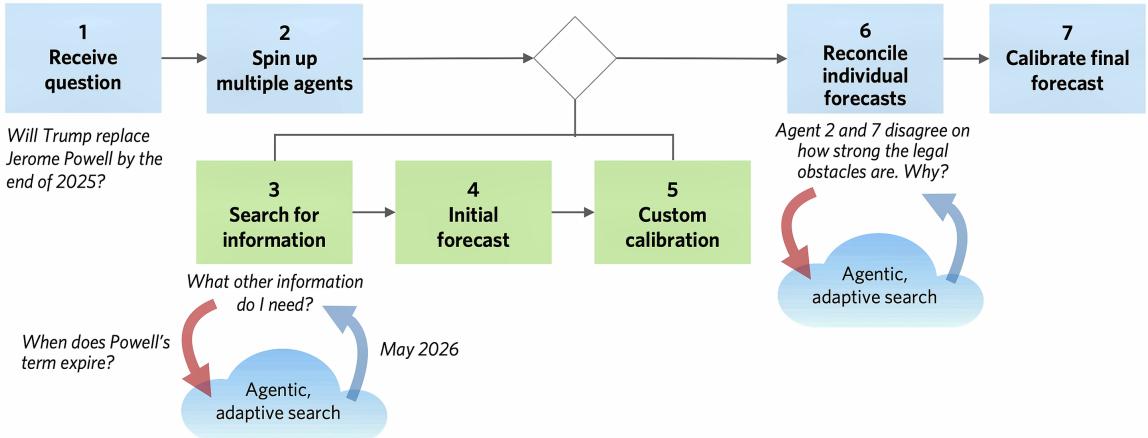


Figure 1: The architecture of the AIA FORECASTER

We now describe the individual components of the AIA FORECASTER.

Search. Each forecasting agent is granted access to an external search provider, which it can query as needed. As described above, a key feature of this search procedure is that it is *agentic* and *adaptive*: each agent is granted full discretion to determine whether and how to query the search provider, and can iteratively improve the quality of its evidence by conditioning each query on prior results. Each forecasting agent operates independently, and their paths can diverge significantly over the course of a single run. As we discuss below, this can be regarded as both a bug and a feature. We can denote this multi-stage evidence gathering pipeline in mathematical notation as

$$\pi : q \rightarrow \mathcal{E}_1 \rightarrow \mathcal{E}_2 \rightarrow \dots \rightarrow \mathcal{E}_n \rightarrow p$$

While agentic search produces far better results than the non-agentic search considered in prior work (Halawi et al., 2024; Karger et al., 2024), this change is insufficient to achieve superforecasting levels. There are multiple reasons for this, all of which can be distilled into the fact that individual forecast runs are unstable, and this instability effectively destroys the fidelity of aggregated benchmark results. The instability itself comes from many sources: missing a key piece of information, over-anchoring to an incorrect base rate, and the inherent randomness of LLMs leading to the occasional outlier opinion.

Reconciliation. One naive solution for addressing run-over-run instability is to simply execute forecasts many times from scratch and to ensemble the results. We show that taking a naive mean over N forecasts is a very strong baseline, and that many advanced ensembling methods fail to outperform this comparatively simple strategy.

In particular, suppose we try to beat the mean by asking an LLM to carefully read all the individual forecasts, evaluate their relative merits, and then produce a final estimate (as proposed in Halawi et al. (2024)). When we do this, we see somewhat surprisingly that the resulting aggregated prediction is substantially less accurate than a simple average of predictions. By inspecting the reasoning traces, we find that LLMs are prone to overemphasizing outlier opinions. This failure exacerbates the instability inherent to LLM forecasting, putting us in a difficult position.

However, it turns out that a simple reframing of this problem is enough to turn our fortunes around. Rather than asking a supervisor LLM to aggregate all forecasts directly, we instead instruct a supervisor agent to examine all forecasts’ reasoning traces \mathcal{R}_i and identify sources of ambiguity or disagreement. This supervisor is allowed to output N search queries, with the hopes that these will help reconcile the differences between competing forecasts. Useful queries at this stage often involve looking up a base rate, or fact-checking assertions made by individual forecasting agents.

Thus, if we denote information gathering agent i as π_i , our fully realized evidence gathering pipeline can be written as

$$\begin{aligned} \pi_i : q &\rightarrow \mathcal{E}_1 \rightarrow \mathcal{E}_2 \rightarrow \cdots \rightarrow \mathcal{E}_n \rightarrow (\mathcal{R}_i, p_i) \\ \text{Supervisor} : (\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_M) &\rightarrow \mathcal{E}_{\text{supervisor}} \rightarrow p_{\text{final}} \end{aligned}$$

Hedging. Finally, we have one more piece of drama to address. It turns out that LLMs have an annoying bias, likely stemming from post-training via reinforcement learning from human feedback (RLHF) (Kaufmann et al., 2023), which makes them hedge toward 0.5 when making predictions. Forecasts for events where the outcome is somewhat certain (e.g., a true probability of 0.85) are often attenuated to a more entropic forecast (e.g., 0.6). Additionally, even for questions where the outcome is certain, LLMs will often predict 0.95 rather than 1.0, justifying their response with reasoning like “unexpected things might happen, and we need to be safe.” We document some examples in Appendix G. These

pitfalls can be corrected rather easily with statistical corrections such as Platt scaling. The core idea is to extremize the forecasts (Baron et al., 2014), correcting the output probabilities along a sigmoid curve. In practice, this process pushes low probabilities further down as they get closer to 0.0 and high probabilities higher up as they get closer to 1.0. We explore popular statistical calibration methods, and demonstrate that Platt scaling and a certain class of extremization techniques are mathematically equivalent. Overall, we see that Platt scaling is an essential statistical correction technique for achieving the best possible performance.³

We now turn to our primary results.

4 The AIA Forecaster Is Indistinguishable from Superforecasters

4.1 Benchmarks and Evaluation Metrics

We report the performance of the AIA FORECASTER on three variants of the FORECASTBENCH benchmark (Karger et al., 2024), as well as an additional benchmark (the MARKETLIQUID benchmark), which we describe below. Each of these benchmarks covers a diverse set of binary questions spanning macroeconomics, politics, technological developments, sports, and other miscellaneous events. For example, a representative question in FORECASTBENCH asks:

Will an infrastructure disaster costing >\$1B in a G20 country be widely attributed to an AI cyberattack before 2025?

Each event in these benchmarks resolves between July 2024 and June 2025, allowing us to observe the ground truth outcome (whether or not the event occurred).

Name	Questions	Resolution Range	Description
FB-7-21	498	7/28/2024 – 12/31/2050	Original Karger et al. (2024) benchmark.
FB-8-14	602	7/28/2024 – 12/31/2050	Updated Karger et al. (2024) benchmark.
FB-MARKET	76	7/25/2025 – 12/31/2050	Subset of Karger et al. (2024) sourced from prediction markets.
MARKETLIQUID	1610	4/2/2025 – 5/23/2025	More challenging benchmark sourced from liquid prediction markets.

Table 1: Overview of benchmarks used to evaluate the performance of the AIA FORECASTER.

ForecastBench. The first variant of FORECASTBENCH, which we call FB-7-21, is the 498-question benchmark previously studied in Karger et al. (2024). The authors report performance for a wide variety of LLM-based forecasters on this benchmark, as well as

3. Interestingly, we consistently found this class of mathematical techniques to be more robust than simple prompting changes, which we found largely ineffective at mitigating this problem.

the aggregate performance of human-generated forecasts in both general public surveys and surveys of expert superforecasters.⁴ These questions are collected from a variety of public prediction markets and other established datasets which track world events. For additional detail on the data collection process we refer to Karger et al. (2024).

We also separately report performance on two variants of **FORECASTBENCH**. The first, **FB-MARKET**, is the subset of **FB-7-21** which is sourced from public prediction markets.⁵ The second, which we call **FB-8-14**, is an updated snapshot of **FORECASTBENCH** as of 08/14/2025 containing 602 questions.

MarketLiquid. Next, we curate a timely dataset of binary forecasting events sourced from a popular public prediction market platform. Our goal was to create a dataset such that (1) all of the markets resolve beyond the knowledge cutoff for the models considered in this manuscript and (2) includes events which are more relevant for policymakers and executives. For example, we restrict our attention to markets covering politics and policy, economics and financial markets, and developments in AI and technology; we exclude markets related to sports, weather prediction, and other miscellaneous categories which appear frequently in **FORECASTBENCH**.

This dataset includes all markets which resolved between 4/2/2025 and 5/23/2025, but excludes markets which (1) were open for less than one week and (2) had fewer than 5,000 total contracts traded. This filtering yields 322 forecasting questions. For each question, we generate binary event forecasting questions at five dates spaced between the posted market open and market resolution date for a total of 1610 questions.⁶ We provide additional detail on the curation of this benchmark, including the choice of dates and the prompts used to convert the raw data into a structured benchmark, in Appendix F.

Evaluation criteria. As is standard in the LLM forecasting literature (e.g., Halawi et al. (2024); Karger et al. (2024); Schoenegger et al. (2024)), we use the Brier score to evaluate forecasting performance (Brier, 1950) on each of the benchmarks described above. The Brier score for an individual forecast $p \in [0, 1]$ with ground truth outcome $o \in \{0, 1\}$ is simply the squared difference of the two, $(p - o)^2$. The Brier score for a set of events indexed by $i = 1, \dots, n$ is the mean of this quantity, $\frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2$. The Brier score ranges from 0 (perfect accuracy) to 1; a forecaster who uniformly predicts $p = 0.5$ for every event can achieve a baseline score of 0.25. The Brier score is known to be a strictly

-
4. Karger et al. (2024) also publish a larger “LLM-only” benchmark, which includes synthetically generated conjunctions of the original questions. We focus here on the benchmark which includes human performance.
 5. As shown in Table 1, the **FB-MARKET** dataset includes questions with resolution dates as late as December 31, 2050. Additionally, a few questions are event-contingent and resolve immediately when the specified event occurs, provided it happens before a designated cutoff date (e.g., year 2100). For these questions, the resolution dates are indeed set at a far future date. Thus, the data set accommodates both fixed-date predictions and open-ended event forecasts.
 6. Note that the market close date (the posted date by which the market will resolve) and the realized resolution date may be distinct, e.g., for events of the form “Will X happen by Y closing date” that occur strictly before the closing date. While we generate questions uniformly between the market open and *realized* resolution date, we only disclose the originally posted close date to the LLM forecaster.

proper scoring rule, which incentivizes truthful forecasting (see e.g., Gneiting and Raftery (2007) for an overview).

4.2 Forecaster Performance

In this section we present our headline results on FB-MARKET, FB-7-21, FB-8-14 and the MARKETLIQUID benchmarks. For FB-7-21, we include the performance of the following three baselines published in Karger et al. (2024): the state-of-the-art LLM forecaster on the Karger et al. (2024) leaderboard, the median forecast for each event in a general public survey, and the median forecast for each event in a survey of expert superforecasters.⁷ We also include the performance of OpenAI’s o3 reasoning model, which we find is the best off-the-shelf forecaster for which these questions do not appear in its pre-training corpus (see Appendix B for details). On FB-MARKET, we further evaluate performance against the market consensus (see Section 5 for details on prediction market mechanics). For FB-8-14, we again include the performance of OpenAI o3, the median of public surveys and superforecasters, and the best-performing LLM forecaster on FORECASTBENCH. For MARKETLIQUID, we include the performance of the prediction market price as a baseline. These results are presented in Table 2 below.

Forecaster	FB-MARKET	FB-7-21	FB-8-14	MARKETLIQUID
Market	0.0965	×	×	0.1106
Public Survey (median per-event)	0.1035	0.1451	0.1510	×
Superforecasters (median per-event)	0.0740	0.1110	0.1152	×
FORECASTBENCH state of the art	0.107	0.133	0.145	×
OpenAI o3	0.1096	0.1221	0.1262	0.1324
AIA FORECASTER	0.0753	0.1076	0.1099	0.1258

Table 2: The performance of the AIA FORECASTER on FORECASTBENCH and MARKETLIQUID. The AIA FORECASTER is statistically indistinguishable from expert superforecasters. Uncertainty quantification is provided in Appendix A.

In contrast to prior work, which shows that LLM forecasters dramatically underperform superforecasters, the AIA FORECASTER is *statistically indistinguishable from superforecasters* on FORECASTBENCH. Indeed, the point estimate for our system’s performance is lower (i.e., more accurate) than the superforecaster baseline. The AIA FORECASTER also substantially outperforms the o3 baseline, the general public survey, and the state-of-the-art LLM-based forecaster on the FORECASTBENCH leaderboard.

7. We follow Karger et al. (2024) in using the median forecast for each event as the corresponding baseline for both superforecasters and the general public survey. This choice can be motivated by arguments which demonstrate that, under the assumption that individual forecasters make symmetric errors around the “true” forecast, the median is an unbiased estimate of this unknown value but the mean is not (Baron et al., 2014). Our empirical evidence corroborates this intuition: the per-event median of superforecasters (0.1110) modestly outperforms the per-event mean (0.1123), and the per-event median of the general public survey (0.1450) substantially outperforms the per-event mean (0.1586). The general public survey averages 49 responses per question (with a minimum of 40). The survey of superforecasters includes an average of eight responses per question (with a minimum of three).

However, FORECASTBENCH is a relatively easy forecasting benchmark. On the more challenging MARKETLIQUID benchmark, the AIA FORECASTER trails the performance of the market consensus. However, as we discuss in Section 5.3, the AIA FORECASTER is *diversifying* to this consensus, and an ensemble of the two outperforms either the market consensus or the AIA FORECASTER considered separately. We now turn to arguably the most important component of a high-quality forecasting system: the search pipeline.

5 Understanding Search in Forecasting

Search is a critical component of human forecasting, as timely news affects the likelihood of an event occurring. Efficient markets, by definition, aggregate salient information into market prices (Fama, 1970). We can see this effect in public prediction markets, which offer contracts that pay a fixed amount if an event occurs (and, typically, a separate contract which pays a fixed amount if it doesn't occur). This allows us to directly observe the market-implied odds of an event occurring.⁸ For example, consider the question: will Grok 4 be released before the end of 2025? On June 27, 2025, Elon Musk announced that Grok 4 would be released just after July 4, 2025. This news announcement caused the market-implied odds to jump to 95-99%, with the market resolving to yes on July 22. We show the market price movement in Figure 2.

However, prior work for LLM forecasters suggests that search does not affect LLM forecasting performance! For example, Karger et al. (2024) find that the best-performing LLM-based forecaster omits search entirely, and adding a search pipeline actually *reduces* forecasting accuracy. Schoenegger et al. (2024) similarly do not find evidence that access to internet search improves forecasting performance.

In this section, we show that advanced forms of search are indeed critical for performance on a wide range of markets and that the structure of prior LLM forecasting benchmarks strongly affects the ability of search to help LLM forecasters.

To do so, we must first deal with the challenge of *foreknowledge bias*. All of the benchmarks described in Section 4 consist entirely of events which have already occurred, which enables us to evaluate performance by comparing forecasts to ground truth outcomes. Of course, this evaluation is rendered moot if we are unable to simulate how the forecaster would have behaved in the past, by blinding it to e.g., contemporaneous news articles which describe the event in question.⁹ We show that this is a surprisingly challenging problem and detail our mitigation strategies.

-
- 8. We use “market price” and “market-implied odds” interchangeably, ignoring transaction costs, margin requirements, and other aspects of market structure. A trader seeking to profit from informed forecasts would of course not have this luxury.
 - 9. Sarkar and Vafa (2024) describe foreknowledge acquired via pre-training as “lookahead bias.” We use the term foreknowledge bias to include other forms of foreknowledge, including information that is exposed via search results.

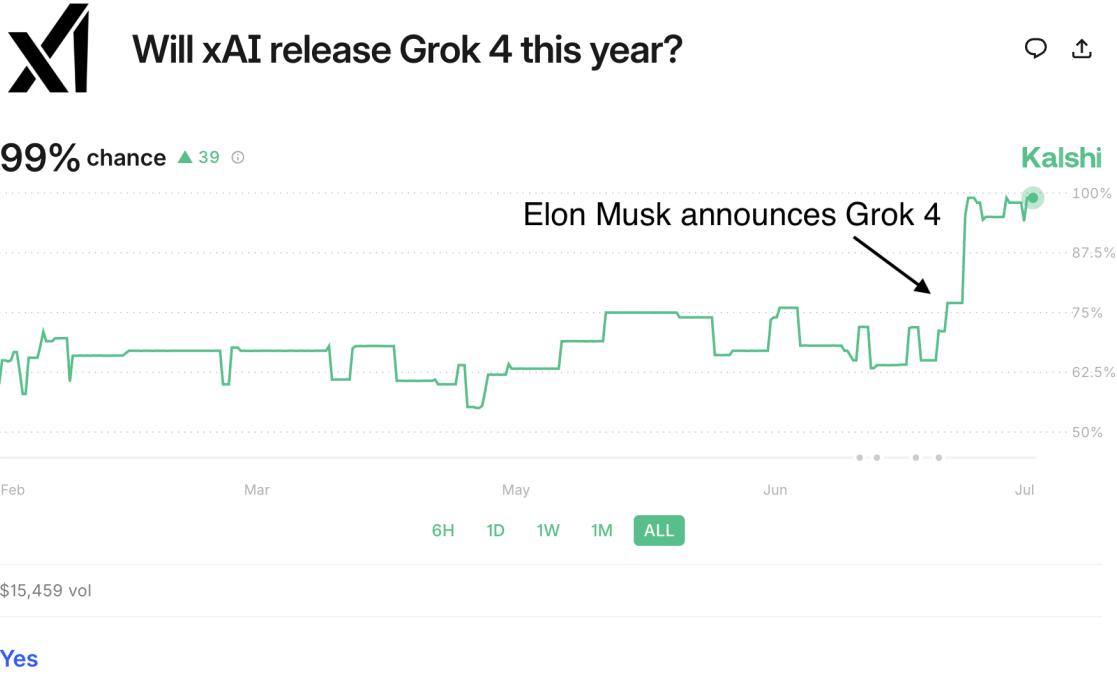


Figure 2: Example of news affecting the market price of a prediction market.

5.1 Search Is Critical for Forecasting

In this section, we quantify the impact of search on forecasting performance. As discussed above, timely information is critical for forecasting, and breaking news events can drive large movements in crowd forecasts. Furthermore, this timely information can only be incorporated in LLM-based forecasters via external tools, as the underlying models are retrained infrequently. This induces so-called “knowledge cutoffs,” at which point the LLM’s knowledge is effectively frozen. For example, OpenAI reports a knowledge cutoff of May 31, 2024 for its flagship o3 reasoning model (OpenAI, 2025), which means the base model has no knowledge of events occurring after this date.¹⁰

We show that high-quality search dramatically improves forecasting performance, and that prior results to the contrary are partially explained by relatively simple, non-adaptive search pipelines. Second, we show that this effect is strongly mitigated by the practice, which is common in the literature, of directly providing prediction market prices in forecasting prompts. These market prices effectively summarize large volumes of relevant information, and can thus compensate for naive or no-search pipelines.

Agentic search improves forecasting performance. We first conduct a series of experiments showing that our agentic search pipeline outperforms a no-search baseline. In order to isolate the effect of search on forecaster performance, we omit the consistency analysis and Platt scaling. We further evaluate two industrial-scale search providers, which

¹⁰. The notion of a single “knowledge cutoff” is a simplification (Cheng et al., 2024), but nonetheless is a good heuristic for understanding where the model’s general knowledge ends.

we anonymize as SEARCH-A and SEARCH-B. While both providers enable general-purpose semantic search over the internet, SEARCH-A focuses primarily on serving natural language user queries, while SEARCH-B is intended to integrate directly with agentic AI systems. We query both providers via their APIs using both “agentic” and “non-agentic” search pipelines. In the former, the LLM-based forecaster can independently define and execute a query strategy, asking follow-up questions as needed. In the latter, the researcher defines the search pipeline, first prompting the forecaster to generate a handful (here, three) search queries and then providing the results as context to subsequent LLM generations. This latter non-agentic strategy is similar to the one used in Halawi et al. (2024), Karger et al. (2024), and Tao et al. (2025).

As shown in Table 3, our agentic search pipeline with access to the SEARCH-A API and our internal corpus of high-quality news data significantly outperforms a no-news baseline on FORECASTBENCH (0.11435 to 0.1230).

Search Method	Brier score
None	0.1230
SEARCH-B (non-agentic)	0.12168
SEARCH-B (agentic)	0.11824
SEARCH-A (non-agentic)	0.11738
SEARCH-A (agentic)	0.1140

Table 3: A comparison of different search strategies. Agentic search outperforms non-agentic search, which in turn only modestly outperforms a no-search baseline.

We also consistently find that, conditional on the source of news, agentic search significantly outperforms non-agentic search, and conditional on the search methodology, SEARCH-A consistently outperforms SEARCH-B. We conjecture that this is due to the relatively low density of relevant information in the typical web page returned by SEARCH-B, while SEARCH-A offers direct responses to questions that are narrowly tailored to the event in question.

We also find that the weakest search method (non-agentic queries to SEARCH-B) performs comparably to simply omitting the retrieval pipeline entirely on FORECASTBENCH. This is consistent with Karger et al. (2024) and Schoenegger et al. (2024), which find no evidence that search improves forecasting performance. However, when viewed in light of our results, this simply highlights the fact that effective search is a nontrivial problem.

The impact of timely news is mediated by market prices. We now show that the effects of search are strongly mitigated by providing the forecaster with access to market prices. This fact provides a partial explanation for prior work which suggests that search does not meaningfully impact forecasting performance (Karger et al., 2024; Schoenegger et al., 2024).

We conduct a series of experiments in which we execute our forecaster on the cross product of: with and without search, and with and without market prices on FORECASTBENCH. We present these results in Table 4.

We find that simply providing the forecaster with prediction market prices improves the performance of a no-news baseline from 0.116 to 0.103 on the subset of FORECASTBENCH derived from public prediction markets—an 11.2% improvement. Providing the market price to our best-performing configuration improves performance from 0.085 to 0.075 on the same subset, a similar 11.8% improvement. In these experiments, we prevent the AIA FORECASTER from retrieving prediction market prices unless they are explicitly provided.¹¹

Includes Market Price?	Forecasting Method	Brier score
✗	LLM, no search	0.116
✓	LLM, no search	0.103
✓	Market price only	0.096
✗	LLM, agentic search	0.085
✓	LLM, agentic search	0.075

Table 4: The interaction of market prices and search. Providing the LLM with access to market prices closes $\sim 42\%$ of the gap between the no-search and agentic search baselines. This suggests that the impact of search is strongly mitigated by access to market prices, which effectively summarize large volumes of relevant information.

We contrast these to the results of Karger et al. (2024), who find that the best-performing LLM-based forecaster also requires access to market prices and achieves a Brier score of 0.109 on this subset.¹² However, the market price alone achieves a Brier score of 0.096 on the same subset of questions; that is, the best-performing LLM-based forecaster in Karger et al. (2024) would perform better if it simply *output the market price which is provided in its prompt*. The AIA FORECASTER also benefits from access to the market price, but performs significantly better than using the market price alone (0.096 for the market price versus 0.075 for the AIA FORECASTER).

Taken together, these results suggest caution in interpreting the performance of LLM-based forecasting systems which have direct access to prediction market prices. On the one hand, these markets can be highly efficient information aggregators, and human and artificial forecasters can both benefit from incorporating these prices into their own predictions. Prediction markets also serve as a convenient testbed for evaluating forecasting performance. On the other hand, access to market prices can mask important differences in forecasting capability, particularly when extrapolating performance to domains where no such prices are available.

11. Specifically, we blacklist domains associated with the largest public prediction markets.

12. Karger et al. (2024) report that the best overall forecaster achieves a Brier score of 0.109 on this subset. The model which performs best on this subset of questions (but worse on the full benchmark) achieves a score of 0.096 on this subset.

5.2 Mitigating Foreknowledge Bias in Search

We now turn to the problem of foreknowledge bias. As discussed above, evaluating search-augmented forecasting systems presents a fundamental challenge: search providers can inadvertently leak information about events that have already occurred. In the extreme case, this bias can create the illusion of forecasting skill where none exists.

For example, consider evaluating an LLM forecaster on the following event: “Will Alphabet’s stock price reach \$190 by December 2024?” The resolution for this event is already known as of this writing (Alphabet stock crossed the \$190 threshold in December 2024), which allows us to evaluate forecasting skill by, e.g., blinding the LLM to events occurring after July 2024 and asking it to provide a forecast. Unfortunately, if the forecaster generates a benign-sounding search query like “Google stock price analysis”, the search provider may return an article simply stating the factual outcome of this event. This can occur even if the search provider is instructed not to return any articles published after July 2024 — as it turns out, determining the date when an article was published (and, perhaps, subsequently updated) is a challenging problem.

In this section, we estimate the prevalence of foreknowledge bias and bound its effect on our results. Through manual analysis, we taxonomize sources of foreknowledge bias and provide examples in Appendix C. We then develop an LLM-as-a-judge pipeline that reviews search provider responses and flags those that exhibit foreknowledge bias.

For each query, we collect search provider snippets and any passages cited by the forecaster, prompt the judge to indicate whether the evidence implies knowledge beyond the intended date cutoff, and record a binary flag with a brief rationale. The exact prompt and JSON schema appear in Appendix D.

Validation of the judge. We manually audit $n=502$ search provider traces randomly sampled across both flagged and unflagged items. Table 5 summarizes performance. As intended for a high-recall detector, flags are noisy positives (low precision), while “clean” classifications are highly reliable (low false negative rate). Practically, this makes it relatively safe to trust unflagged runs, while treating raw flag counts as an overestimate of true foreknowledge bias.

Classification (n=502)	Count	Percentage
True Positives (correctly identified leaks)	16	3.2%
False Positives (flagged clean responses)	65	12.9%
True Negatives (correctly identified clean)	412	82.1%
False Negatives (missed leaks)	9	1.8%

Table 5: Validation of the judge on 502 randomly sampled traces. While the flags are noisy (a relatively high false positive rate), responses which are not flagged reliably exclude foreknowledge (low false negative rate).

From this exercise, we estimate that $\sim 1.65\%$ of all search results include some form of foreknowledge bias. We provide additional details for this calculation in Appendix E. With this estimate in hand, we now seek to bound the impact of foreknowledge bias on forecasting performance.

Robustness checks. We test whether residual foreknowledge bias affects our performance metrics through two conservative robustness checks:

- **Filtered (all flags):** Remove every judge-flagged response. Roughly 80% of removed forecasts do not exhibit foreknowledge bias, making this test intentionally conservative.
- **Worst-case:** Assume a forecast of 0.50 (corresponding to a Brier score of 0.25) for any question with ≥ 5 flags. This is again intentionally conservative, assuming no forecasting skill on any questions which appear to suffer from foreknowledge bias.

We report performance without the consistency analysis and statistical correction steps to isolate the effect of foreknowledge bias.

Analysis	Brier score	Change
Baseline	0.1159	—
Filtered	0.1161	+0.17%
Worst-case	0.1166	+0.55%

Table 6: Robustness checks to evaluate foreknowledge bias on FORECASTBENCH. “Filtered” computes the Brier score after removing all questions flagged by the judge. “Worst-case” assumes a forecast of 0.5 for all flagged questions. Both perform comparably to a no-mitigation baseline.

As Table 6 demonstrates, both robustness checks yield a Brier score that is within 0.001 of the baseline value ($\leq 0.6\%$ on a relative basis). Thus, we conclude that foreknowledge bias does not appear to materially affect performance.

5.3 LLM Forecasts Are Diversifying to Market Prices

We now investigate the complementarity between LLM forecasts and market prices. In Section 4 we saw that our forecaster outperforms the market consensus on FORECASTBENCH, but modestly underperforms on the stronger MARKETLIQUID benchmark. It is natural to then ask: in cases where the forecaster outperforms the market, does the market price provide any *additional* predictive signal? And conversely, in settings where the forecaster underperforms the market consensus, can we disregard its forecasts entirely?

We investigate these questions by estimating a bivariate simplex-constrained regression of the binary resolution indicator on two covariates: the market price and the LLM forecast. This regression learns a convex combination of the two covariates that minimizes the

Brier score (i.e., the standard ℓ_2 loss). Our goal is not to produce the best possible ensemble forecast, but rather to derive interpretable coefficients that serve as proxies for the relative information each forecast provides about the event of interest. We report point estimates for these coefficients along with confidence intervals computed from 1000 bootstrap samples. To contextualize these results, we include the Brier score achieved by a leave-one-out version of this simplex regression procedure, providing an out-of-sample estimate of the ensemble’s performance.

Benchmark	AIA Forecaster		Market		Ensemble Brier score
	Brier score	Coefficient	Brier score	Coefficient	
FB-MARKET	0.075	0.87 [0.42, 1.00]	0.097	0.14 [0.00, 0.58]	0.079
MARKETLIQUID	0.126	0.33 [0.12, 0.47]	0.111	0.67 [0.53, 0.80]	0.106

Table 7: The performance of a simplex-constrained regression ensemble which combines LLM forecasts with the market price. Regression coefficients are reported with 95% bootstrap confidence intervals over 1000 samples. Ensemble performance is reported as a leave-one-out estimate. Coefficients may not sum to 1 due to rounding.

As Table 7 demonstrates, the AIA FORECASTER is assigned the majority of the ensemble weight on the FB-MARKET benchmark, and the leave-one-out ensemble performs worse than the AIA FORECASTER alone. Thus, the market price does *not* provide signal that is additive to the forecasts produced by our system. However, on MARKETLIQUID, where the market price outperforms the AIA FORECASTER, the converse is not true. Instead, the ensemble assigns roughly a third of its weight to the LLM forecasts, and the leave-one-out ensemble outperforms both the LLM forecasts and the market price. This suggests that even in challenging settings where LLMs fail to outperform the market consensus outright, they can still provide valuable and diversifying information to decision makers.

6 Search Is Critical for Live Markets

We now turn to evaluating the AIA FORECASTER on prediction markets which have not yet resolved (“live markets”) at the time of forecaster execution. This kind of evaluation precludes any form of foreknowledge bias, as we are asking the model to predict events which have not yet occurred.

To do this, we collected a dataset of live markets on a popular prediction market platform, filtered by relevance and liquidity. Our filtering process for the live markets was the same process used to generate the static MARKETLIQUID dataset (except filtered to markets that are currently open). We conducted this process nightly between August 15, 2025 and August 21, 2025, which yielded approximately 1500 questions per night. We sampled to 250 questions at random every night to generate the evaluation set. These questions changed on a nightly basis as new markets opened and existing markets were resolved.

We conduct two sets of experiments. For both experiments, we executed the AIA FORECASTER with and without search enabled. In both conditions, we omitted the market prices to isolate the effect of search on forecaster performance.¹³

Market Type	With Search (Brier score)	Without Search (Brier score)	Market Baseline (Brier score)
Closed markets (n=64)	0.1002	0.3609	0.1111
Open markets (n=1750)	0.0522	0.0910	—

Table 8: Real-time forecasting performance with and without search. In both cases the system is prevented from accessing market prices. In closed markets, forecasts are evaluated against ground truth outcomes. In open markets, forecasts are scored against market prices. The performance of the market consensus on closed markets is included for reference. AIA FORECASTER improves on the no-search baseline by $3.6\times$ in closed markets and is more concordant with market prices in open markets.

In the first experiment, we evaluate the Brier score on the 64 markets that closed by August 26, 2025. As shown in Table 8, with search, the AIA FORECASTER achieved a Brier score of 0.1002. Without search, the AIA FORECASTER achieved a Brier score of 0.3609, a difference of $3.6\times$. As we can see, search is critical for forecasting; in fact, excluding search yields performance which is worse than uniformly predicting 50% (which mechanically achieves a Brier score of 0.25).

We also compute the Brier score achieved by simply using the market price as a baseline (“consensus”) forecast. We see that, with search, the AIA FORECASTER slightly outperforms the market consensus, although with only 64 markets we are unable to statistically distinguish these two results.

In the second experiment, we evaluate the Brier score *with respect to the market price*. Unlike the Brier score for resolved markets, a perfectly skilled forecaster would score substantially away from 0. Nonetheless, market prices are a strong baseline forecast (Surowiecki, 2004; Bassamboo et al., 2015; Da and Huang, 2019; Lichtendahl Jr et al., 2013), so significant deviations from a market price baseline are unlikely for a skilled forecaster. Importantly, this allows us to evaluate the AIA FORECASTER over the much larger (1750) set of markets which are still open.

As shown in Table 8, compared to market prices, AIA FORECASTER achieves a Brier score of 0.0522 with search enabled. Without search, our forecaster achieves a Brier score of 0.0910. Because we expect market prices to be informative, being close to the market price likely indicates better performance. Thus, we can again see that search is critical for generating informative forecasts.

These experiments on live markets are necessarily narrower in scope than the larger-scale retrospective evaluations we conduct in Section 4. Nonetheless, we again see evidence

13. As in Section 5.1, we exclude market prices from the prompt and prevent the forecaster from accessing domains associated with the most popular prediction market platforms.

which corroborates our main findings, and further suggests that foreknowledge bias is not meaningfully impacting our results.

7 Ensembling, Synthesis, and Statistical Corrections

In this section, we explore various approaches to ensembling multiple forecasts into a stronger aggregate forecast. We find that some form of ensembling is critical and dramatically outperforms naive baselines. In Section 7.1, we motivate the need for ensembling forecasts, and find that simple averaging of independent forecasts (and minor variations thereof) dramatically outperforms a naive baseline which eschews any form of aggregation. In Section 7.2, we find that an additional “supervisor” agent can substantially improve on the performance of simple mean aggregation by actively resolving disagreements between individual forecasts. Finally, in Section 7.3, we find that the additional application of Platt scaling improves our performance to match that of superforecasters.

7.1 Simple Averaging Is a Strong Baseline

We now explore the performance of simple aggregation methods. First, to motivate the need for aggregating multiple forecasts, consider the following simple model. Let the random variables $Q \in \mathcal{Q}, P \in [0, 1], O \in \{0, 1\}$ denote questions, forecasts and outcomes, respectively. Let (Q, P, O) be a random vector with some well-defined probability measure. Let $\ell(P, O) = (O - P)^2$ be the Brier score. Under this model, fixing a question $Q = q$ induces some conditional distribution over the question-specific Brier score ℓ . Because ℓ is strongly convex, the following is an immediate consequence of Jensen’s inequality:

$$\ell(\mathbb{E}[P | Q = q], \mathbb{E}[O | Q = q]) < \mathbb{E}[\ell(P, O) | Q = q] \quad \forall q. \quad (1)$$

That is, the Brier score for the average forecast and outcome is a strict improvement on the Brier score averaged over possible forecasts and outcomes.¹⁴ For example, given the choice between (1) predicting exactly 50% or (2) predicting 100% and 0% with equal probability, the former is preferable *regardless of the true outcome*. This suggests that averaging many independent forecasts for a given question will tend to improve on sampling a single forecast. We confirm this empirically in Figure 3 below, which demonstrates sharply decreasing Brier scores on MARKETLIQUID as we increase the number of independent forecasts from 1 to 5, and an additional modest improvement as we further increase the number to 15 forecasts.¹⁵

14. Breiman (1996) uses this fact to motivate ensembling in the classical statistical learning paradigm.

15. To maintain computational tractability, Figure 3 presents the results of bootstrap resampling a fixed number of forecasts from superpopulation of 50 forecasts per question rather than repeatedly re-evaluating the benchmark at each point on the x-axis. We use the larger MARKETLIQUID benchmark to increase statistical power.

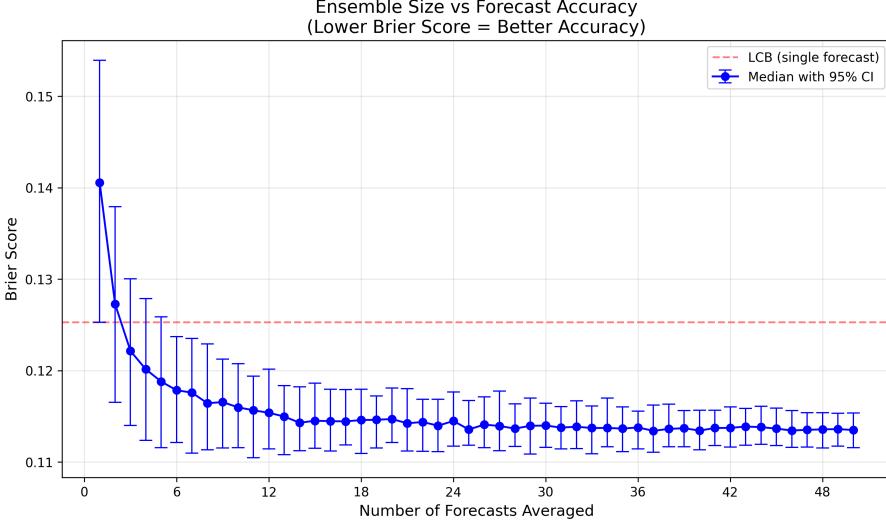


Figure 3: The Brier score induced at various ensemble sizes. Point estimates and 95% confidence intervals are generated via bootstrap resampling from a set of 50 forecasts per question. The dashed line indicates the lower confidence bound for a single forecast.

Synthesis Method	Brier score
Simple Mean	0.1140
Trimmed Mean	0.1142
Median	0.1138
Single Forecast	0.1182

Table 9: Evaluating different approaches to ensembling multiple forecasts on FB-7-21. The mean, trimmed mean and median aggregation over 10 forecasts all perform substantially better than a single randomly chosen forecast.

Figure 3 also suggests diminishing returns beyond a handful of independent forecasts; we adopt 10 forecasts per question as a conservative standard in this work. This result also illustrates that sampling only a single forecast for each question yields highly variable performance—even when averaged over the 1610 questions in this evaluation set—due to the inherent randomness in LLM outputs. This variability is itself an undesirable property for a forecaster, and suggests that some form of ensembling is critical for internal validity. This kind of ensembling has thus far not been common in the LLM forecasting literature¹⁶, and suggests that the common practice of quantifying uncertainty *only over forecasting questions themselves*—rather than also considering uncertainty over LLM outputs—dramatically understates the true variability of LLM forecasting performance.

In Table 9, we further demonstrate that minor variations of this approach, like taking the median of 10 forecasts, or taking the trimmed mean (by discarding the smallest and largest of the 10 forecasts), perform comparably to simple averaging. We include a reference baseline which samples a single individual forecast for each question uniformly at random.

16. One exception is Schoenegger et al. (2024), though this work is motivated by the hypothesis that ensembling *different* LLM-based forecasters will improve performance.

The results above suggest that averaging multiple independent forecasts is critical for forecasting performance and cannot be meaningfully improved by minor variations thereof.

7.2 Supervisor Agents Resolve Disagreements

We now turn to more sophisticated ensembling methods, which utilize an LLM-based supervisor to synthesize multiple independent forecasts. This supervisor can be viewed as a kind of generative verifier or LLM-as-a-judge (Li et al., 2024; Zhang et al., 2024) which aggregates the 10 forecasts generated by each agent into a final probability.

We consider three different supervisor agents, as well as the same no aggregation baseline described in Section 7.1. The first, which we call the “best of k ” supervisor, is provided with 10 forecasts and asked to reason over them to output the best of these forecasts. Importantly, the supervisor is explicitly constrained to output exactly one of the provided forecasts. The second, which we call the “non-agentic supervisor,” uses the method proposed in Halawi et al. (2024) to ensemble multiple forecasts into a final forecast. Unlike the best of k method, this method is not constrained to output one of the input forecasts. Instead, the model simply uses these as inputs to reason toward a stronger forecast. Finally, we propose the “agentic supervisor,” which proceeds in three steps. First, the supervisor model identifies disagreements among the 10 input forecasts. Second, the supervisor generates and executes a series of search queries intended to resolve these disagreements (e.g., to resolve disagreement over the accurate base rate for an event). Finally, the supervisor generates an updated forecast along with its confidence that the forecast was updated in the correct direction (“high,” “medium” or “low”). High confidence updates are used in place of the simple mean, while medium or low confidence updates are discarded.

Synthesis Method	Brier score	Top@3	Worst@3	Outperform (%)
Agentic Supervisor	0.1125	33.9%	4.6%	3.8%
Non-agentic Supervisor	0.1168	7.8%	6.2%	0.6%
Best of k (LLM)	0.1191	29.7%	7.2%	0.0%
None	0.1199			

Table 10: Comparison of different synthesis strategies on FB-7-21. **Top@3**: fraction of aggregated forecasts which are at least as accurate as the three best individual forecasts; **Worst@3**: fraction of aggregated forecasts which are no more accurate than the three worst individual forecasts; **Outperform (%)**: percentage of aggregate forecasts beating all individual forecasts. Agentic supervision yields the best overall performance.

As shown in Table 10, the agentic supervisor achieves the best performance among the synthesis methods. The best of k approach, which selects a single forecast from the ensemble to serve as the final prediction, performs poorly. It never selects a forecast that is better than any of the candidate forecasts by design, and it happens to pick among the worst of the candidate forecasts with a striking 7.2% frequency. The non-agentic supervisor, in which a language model performs the aggregation, shows modest improvement over

the best of k method. However, this method produces quality forecasts (top@3 of 7.8%) relatively rarely. In contrast, the agentic supervisor substantially outperforms both alternatives. It actively reconciles inconsistencies between individual predictions and performs additional news queries to refine forecast rationales. This synthesis process enables it to achieve the lowest Brier score of 0.1125, highest top@3 accuracy, highest outperformance rate, and lowest worst@3 rate, demonstrating the value of additional search and agentic reasoning when aggregating forecasts.

7.3 Statistical Corrections and Where They Help

Statistical corrections are commonly applied to forecasts to correct for distortions in the forecasted probabilities ([Niculescu-Mizil and Caruana, 2005](#); [Baron et al., 2014](#)). In the binary case, [Niculescu-Mizil and Caruana](#) note that forecasts often result in probabilities that are clustered away from zero or one, which we observe in our data (Figure 4).

We consider statistical correction functions $f_\theta(p) : [0, 1] \rightarrow [0, 1]$ ([Neyman and Roughgarden, 2022](#)), which may either have learnable parameters ([Niculescu-Mizil and Caruana, 2005](#)) or are deterministic functions ([Satopää et al., 2014](#)).

Correction Method	Fixed Parameter	ID Training	OOD Training
Platt Scaling	0.1076	0.1071	0.1104
Log Odds Extremization	0.1085	×	×
Isotonic Regression	×	0.1097	0.1134
OLS	×	0.1119	0.1125
None	0.1140	×	×

Table 11: Evaluating different statistical correction approaches on FB-7-21. The fixed parameter setting sets the scaling parameter to $\sqrt{3}$, as proposed by [Neyman and Roughgarden \(2022\)](#). In distribution (ID) training reports leave-one-out performance with learned scaling parameters. Out of distribution (OOD) training fits the scaling parameters on the benchmark released by [Halawi et al. \(2024\)](#). Platt scaling yields the best performance across these methods.

Table 11 shows that the most beneficial statistical correction methods are Platt scaling ([Niculescu-Mizil and Caruana, 2005](#)) and extremization ([Neyman and Roughgarden, 2022](#)). Log odds extremization aggregates a set of k forecasts by taking a scaled mean of the log odds of each forecast. The scaling parameter can be learned from data or set to a fixed value a priori. Log odds extremization follows the following formula ([Baron et al., 2014](#)):

$$\log \frac{\hat{p}}{1 - \hat{p}} = \frac{d}{n} \sum_{i=1}^n \log \frac{p_i}{1 - p_i}, \quad (2)$$

where the aggregate estimate \hat{p} can then be recovered using the sigmoid function. We demonstrate that Equation 2 is mathematically identical to applying Platt scaling to the geometric mean of forecasts with a Platt coefficient of d in Appendix G.

Neyman and Roughgarden (2022) propose a parameter value of $d = \sqrt{3}$. We adopt this convention in our work to avoid the risk of overfitting to the benchmarks we consider. Using Platt scaling on the arithmetic mean of k forecasts results in the lowest Brier score, as shown in Table 11.

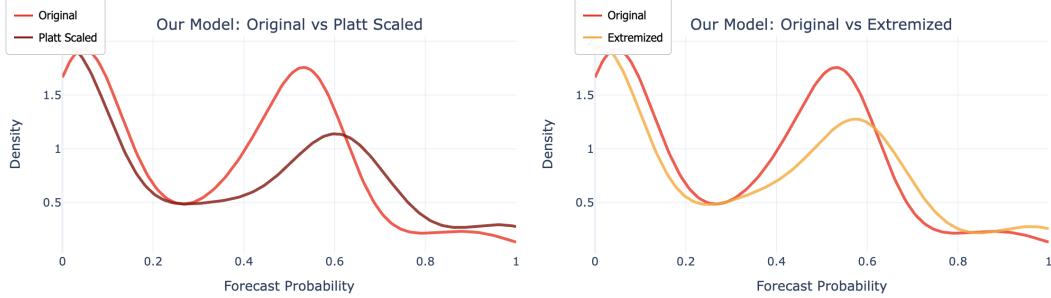


Figure 4: Scaling and extremization shifts the mass of probabilities toward the extremities, especially in the center mass which produces larger drops in Brier scores.

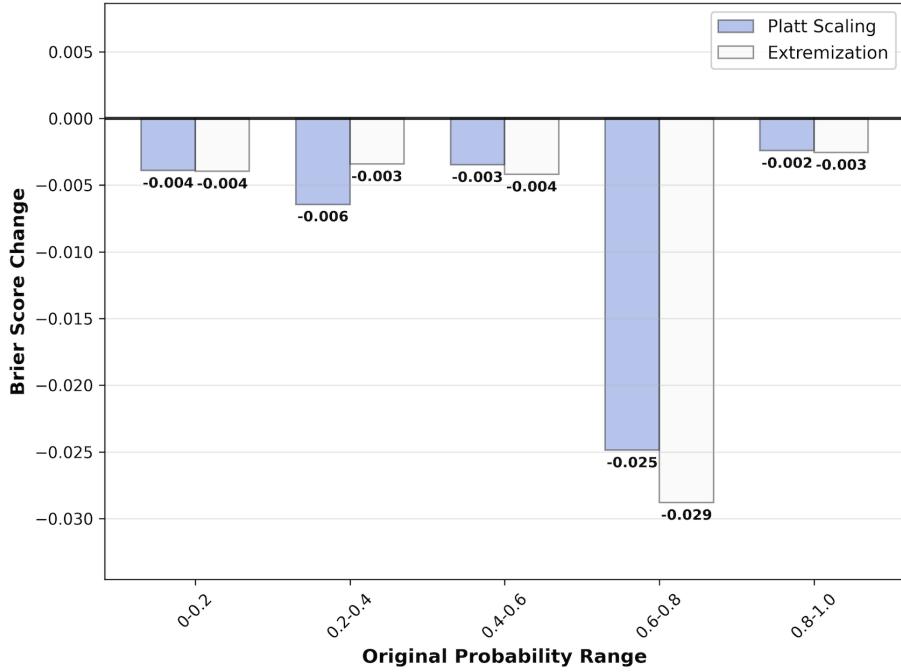


Figure 5: Across probability bins, the largest drops in Brier scores due to correction comes from the 0.6-0.8 forecast bin, followed by the 0.2-0.4 forecast bin.

Mechanisms of Improvement. As Table 11 shows, both Platt scaling and extremization significantly improve forecast accuracy, while isotonic regression and OLS are less competitive. Both Platt scaling and extremization effectively push probabilities toward their respective bounds (0 or 1). This suggests that initial forecast quality is critical to these kinds of statistical corrections, insofar as forecasts on the wrong side of 0.5 would be adjusted in the wrong direction. We also see that the improvement gained by original forecasts that are already confident and correct are (mechanically) limited. For example, an increase in the forecast from 0.995 to 0.999 yields a corresponding change in the Brier score of 2.4×10^{-5} . Meanwhile a change in the forecast from 0.497 to 0.381

yields a corresponding change in the Brier score of -0.102 . That is, Platt scaling and extremization push forecasts towards the extremes. Figure 4 illustrates this redistribution effect, where mass shifts from the $0.5\text{-}0.6$ range to the $0.6\text{-}0.75$ range for events that ultimately resolve to 1.0.

This push toward the tails is what generates the substantial Brier score improvements when less confident but correct predictions become more confident, which we can also observe in Figure 5¹⁷, where the largest Brier score decreases come from original forecasts $p \in [0.2, 0.4) \cup [0.6, 0.8)$.

Although extremization and Platt scaling are both effective and share mathematical similarities, supervisor agents offer significant improvements to forecast quality, as discussed in Section 7.2. The extremization approach detailed in Equation 2 requires aggregating k forecasts, which prevents us from leveraging the supervisor’s contributions. Therefore, we adopt Platt scaling as our correction of choice: it preserves the advantages of extremization while enabling the supervisor resolution step, ultimately yielding more accurate forecasts as demonstrated in Table 11.

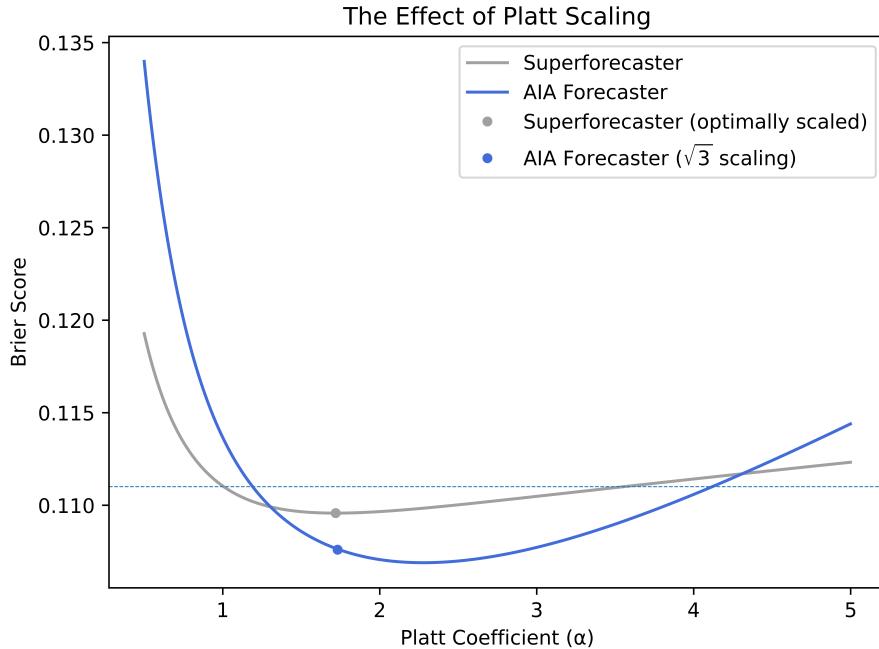


Figure 6: AIA FORECASTER outperforms superforecasters across a range of Platt coefficients, including when the coefficient is optimized for superforecaster performance ($\alpha = 1.72$, gray dot). The AIA FORECASTER uses a coefficient of $\alpha = \sqrt{3} \approx 1.73$, as suggested by Neyman and Roughgarden (2022) (blue dot).

Comparison with Superforecasters. Finally, because statistical corrections are agnostic to the technique used to generate the forecasts, they can also be applied to the superforecaster predictions. Indeed, applying a statistical correction to the superforecaster

17. Near $p \approx 0.5$, Platt scaling and extremization are approximately the identity function. In such scenarios, the brier score difference is minimally impacted as this effect dominates the quadratic Brier score improvement.

predictions is arguably required to produce a fair comparison between human and AI forecasts (Nikos et al., 2024).

In Figure 6, we plot the performance of both superforecasters and the AIA FORECASTER across a range of Platt coefficients. As Figure 6 demonstrates, the AIA FORECASTER outperforms the superforecaster baseline across a wide range of scaling parameters. Furthermore, scaling the AIA FORECASTER with a Platt coefficient of $\sqrt{3} \approx 1.73$, as suggested by Neyman and Roughgarden (2022), outperforms superforecasters even when the scaling parameter is chosen to optimize their performance (a coefficient of 1.72). Optimizing the scaling coefficient for the AIA FORECASTER yields a further modest improvement (at a coefficient of 2.27), though, as described above, we do not use this learned parameter to mitigate the risk of overfitting to a particular benchmark.

8 Discussion

This work describes a practical recipe for LLM-based judgmental forecasting: (i) prioritize evidence quality via agentic, adaptive search; (ii) reduce variance by ensembling independent runs; (iii) use a supervisor agent to actively reconcile disparate forecasts; and (iv) apply statistical calibration to counteract the behavioral biases of LLMs. Together, these components yield performance that is statistically indistinguishable from human superforecasters on FORECASTBENCH, and diversifying to (albeit less accurate than) market consensus in liquid prediction markets.

Methodologically, we find that ensembling is not optional. Single-run performance is both noisier *and worse in expectation* than even simple averaging across a small number of independent generations. This also implies that the practice—which has thus far been common in the LLM forecasting literature—of neglecting to quantify uncertainty of LLM outputs themselves may threaten both internal and external validity of the claimed results. Second, generative verification methods are not necessarily superior to simple averaging. In fact, naive implementations of this approach can perform quite poorly. However, an agentic supervisor that can actively resolve disagreements and seek clarifications yields meaningful gains in forecasting performance. Third, statistical corrections provide a lightweight mechanism for aligning forecasts with scoring rules. However, post hoc calibration only helps when initial judgments are already on the correct side of 0.5, reinforcing that upstream evidence and reasoning are the primary mechanisms for improving forecast quality.

Limitations, risks and future work. A primary theme of this work is the difficulty of evaluating LLM-based forecasters. Perhaps the biggest challenge is the narrow temporal window in which this evaluation can be performed. On the one hand, events occurring before the model’s knowledge cutoff are likely to have been memorized by the LLM (Sarkar and Vafa, 2024), making retrospective analysis difficult. On the other hand, evaluating the LLM on future events is inefficient—we need to wait for the events to occur!—which

precludes rapid iteration. This challenge is compounded by the more general issue of foreknowledge bias in search results (Paleka et al., 2025). LLMs are also known to be prone to hallucination (Huang et al., 2025) and do not always reason in an internally consistent manner (Valmeekam et al., 2023; Vafa et al., 2024). This makes it difficult to extrapolate from performance on one task to other (seemingly) related domains, and we caution that our work does not in any way obviate the need for human oversight in high-stakes decision tasks.

Finally, our work focuses on forecasting *unconditional, binary* events, of the form “will X happen by Y date?” There are many other types of forecasting tasks, including k-class forecasting (e.g., “which equity sector will perform the best over the next year?”), point estimate forecasting (e.g., “what will the next inflation print be?”) and conditional forecasting (e.g., “if the US enters a recession, what will Donald Trump’s approval rating be?”). All of these alternative forms of forecasting are amenable, in principle, to LLM-based forecasting, but will require significant additional work to incorporate into our forecasting system.

Despite these limitations, our work demonstrates the significant promise of LLM-based forecasting. We achieve expert-level performance on the FORECASTBENCH benchmark, and resolve a number of open questions in the literature. We also develop a set of practical, easily portable standards for LLM-based forecasting, including the use of agentic search, ensembling and statistical calibration techniques. Taken together, our findings suggest that AI forecasters may soon achieve superhuman performance.

References

- Jonathan Baron, Barbara A Mellers, Philip E Tetlock, Eric Stone, and Lyle H Ungar. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145, 2014.
- Achal Bassamboo, Ruomeng Cui, and Antonio Moreno. The wisdom of crowds in operations: Forecasting using prediction markets. *SSRN Electronic Journal*, 01 2015. doi: 10.2139/ssrn.2679663.
- Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, August 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. URL <https://doi.org/10.1023/A:1018054314350>.
- Glenn W Brier. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78(1):1–3, January 1950.
- Emanuel Bueechi, Milan Fischer, Laura Crocetti, Miroslav Trnka, Aleš Grlj, Luca Zappa, and Wouter Dorigo. Crop yield anomaly forecasting in the pannonian basin using gradient boosting and its performance in years of severe drought. *Agricultural and Forest Meteorology*, 340:109596, 2023.
- Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*, 2024.
- Alicia Curth, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van Der Schaar. Using machine learning to individualize treatment effect estimation: challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 115(4):710–719, 2024.
- Zhi Da and Xing Huang. Harnessing the wisdom of crowds. *Management Science*, 66, 08 2019. doi: 10.1287/mnsc.2019.3294.
- Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.
- Andrew Gelman, Jessica Hullman, Christopher Wlezien, and George Elliott Morris. Information, incentives, and goals in election forecasts. *Judgment and Decision Making*, 15(5):863–880, 2020.
- Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102(477):359–378, March 2007.
- Google DeepMind. Weathernext: Our most advanced weather forecasting AI technology. 2024. URL <https://deepmind.google/science/weathernext/>.

Google DeepMind. Advanced version of Gemini with Deep Think officially achieves gold medal standard at the International Mathematical Olympiad. July 2025. URL <https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard>

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37:50426–50468, 2024.

Robin D Hanson. Decision markets. 1999.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL <https://doi.org/10.1145/3703155>.

Will Jennings, Michael Lewis-Beck, and Christopher Wlezien. Election forecasting: Too far out? *International Journal of Forecasting*, 36(3):949–962, 2020.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. *arXiv preprint arXiv:2005.00792*, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873): 583–589, 2021.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. 2023.

Kajal Lahiri and Liu Yang. Forecasting binary outcomes. In *Handbook of Economic Forecasting*, volume 2, pages 1025–1106. Elsevier, 2013.

Michael Lawrence, Paul Goodwin, Marcus O’Connor, and Dilek Önkal. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3):493–518, 2006. ISSN 0169-2070. doi: <https://doi.org/10.1016/j.ijforecast.2006.03.007>.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on LLM-based evaluation methods. 2024.

Kenneth Lichtendahl Jr, Yael Grushka-Cockayne, and Phil Pfeifer. The wisdom of competitive crowds. *Operations Research*, 61, 03 2013. doi: 10.2139/ssrn.1926330.

Metaculus. Q1 AI benchmarking results: Pro forecasters crush bots. May 2025. URL <https://www.metaculus.com/notebooks/38673/q1-ai-benchmarking-results/>. Metaculus notebook, published May 2025.

Gerrit Mutschlechner and Adam Jatowt. Analyzing the role of context in forecasting with large language models. *arXiv preprint arXiv:2501.06496*, 2025.

Eric Neyman and Tim Roughgarden. Are you smarter than a random expert? the robust aggregation of substitutable signals. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 990–1012, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538243. URL <https://doi.org/10.1145/3490486.3538243>.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL <https://doi.org/10.1145/1102351.1102430>.

Nikos, Peter Mühlbacher, Lawrence Phillips, and dschwarz. Contra papers claiming superhuman ai forecasting, September 12 2024. URL <https://www.lesswrong.com/posts/uGkRcHqatmPkvpGLq/contra-papers-claiming-superhuman-ai-forecasting>.

OpenAI. OpenAI o3 and o4-mini System Card. System card, OpenAI, April 2025. URL <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>. Accessed: 2025-08-04.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723*, 2025.

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. 2020.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Emelia Probasco, Helen Toner, Matthew Burtell, and Tim G. J. Rudner. AI for military decision-making: Harnessing the advantages and avoiding the risks. Technical report, Center for Security and Emerging Technology, April 2025.

Long Qian, Xin Lu, Parvez Haris, Jianyong Zhu, Shuo Li, and Yingjie Yang. Enhancing clinical trial outcome prediction with artificial intelligence: a systematic review. *Drug Discovery Today*, page 104332, 2025.

Suproteem K Sarkar and Keyon Vafa. Lookahead bias in pretrained language models. Available at SSRN 4754678, 2024.

Ville Satopää, Jonathan Baron, Dean Foster, Barbara Mellers, Philip Tetlock, and Lyle Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30:344–356, 04 2014. doi: 10.1016/j.ijforecast.2013.09.009.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Philipp Schoenegger, Cameron R Jones, Philip E Tetlock, and Barbara Mellers. Prompt engineering large language models’ forecasting capabilities. *arXiv preprint arXiv:2506.01578*, 2025.

Alix Soliman. Deepmind ai weather forecaster beats world-class system. *Nature*, 636: 282–283, 2024. doi: 10.1038/d41586-024-03957-3. URL <https://doi.org/10.1038/d41586-024-03957-3>. Artificial-intelligence model provides forecasts 15 days out, as well as the probability of accuracy. And it does so faster than the best operational model.

James Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. 2004.

Tetsuji Tanaka, Laixiang Sun, Inbal Becker-Reshef, Xiao-Peng Song, and Estefania Puricelli. Satellite forecasting of crop harvest can trigger a cross-hemispheric production response and improve global food security. *Communications Earth & Environment*, 4 (1):334, 2023.

Zhengwei Tao, Zhi Jin, Bincheng Li, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Xiancai Chen, Jia Li, Linyu Li, and Chongyang Tao. Prophet: An inferable future forecasting benchmark with causal intervened likelihood estimation. 2025.

Philip E Tetlock. *Expert political judgment: How good is it? How can we know?-New edition*. Princeton University Press, 2017.

Philip E Tetlock and Dan Gardner. *Superforecasting: The art and science of prediction*. Random House, 2016.

Jorge AV Tohalino and Diego R Amancio. On predicting research grants productivity via machine learning. *Journal of Informetrics*, 16(2):101260, 2022.

Ruey S Tsay. *Analysis of financial time series*. John Wiley & Sons, 2005.

Benjamin Turtel, Danny Franklin, and Philipp Schoenegger. LLMs can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. 2024.

Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. On the planning abilities of large language models (a critical investigation with a proposed benchmark). 2023.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International journal of forecasting*, 31(3):980–991, 2015.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Openforecast: A large-scale open-ended event forecasting dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5273–5294, 2025.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. The future outcome reasoning and confidence assessment benchmark. *arXiv e-prints*, pages arXiv–2502, 2025.

Michael Zequeira. Artificial intelligence as a combat multiplier. *Military Review*, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. 2024.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35: 27293–27305, 2022.

A. Uncertainty Quantification for Primary Results

In this section we quantify uncertainty over the results presented in Section 4. To do this, we follow Karger et al. (2024) in computing pairwise p-values via bootstrap resampling. In particular, we first compute the mean difference in Brier scores at the individual question level, which will serve as our statistic of interest. This is equivalent to simply computing the difference in overall Brier score that two different methods achieve on a given benchmark. We then approximate the null distribution of this statistic by (1) recentering these question-level differences at zero, (2) drawing 10000 bootstrap samples from these recentered differences, and (3) computing the mean difference in Brier score for each of these samples. This procedure will, for example, have power to detect cases where one forecaster performs slightly better than another *uniformly* over all questions, even if the per-question performance gap is small (e.g., the better forecaster always provides a forecast that is 1% closer to the true outcome). In contrast, computing separate confidence intervals for each Brier score, without conditioning on the questions, will struggle to detect this kind of performance gap. Importantly, this procedure does not account for uncertainty over the LLM outputs themselves. We address this separately Section 7. We present these results in Tables 12, 13, 14 and 15 below.

Forecaster	Brier score	Win Rate vs Best	Pairwise p-value
Market	0.0956	0.4212	0.0687
Public Survey (median per-event)	0.1035	0.2236	<0.001
Karger et al. (2024) state of the art	0.107	-	-
OpenAI o3	0.1096	0.6447	0.0305
AIA Forecaster	0.0753	0.3026	0.4328
Superforecasters (median per-event)	0.0740	-	-

Table 12: Uncertainty quantification for results on FB-MARKET in Table 2. Karger et al. (2024) do not publish individual LLM forecasts, precluding a direct pairwise comparison.

Forecaster	Brier score	Win Rate vs Best	Pairwise p-value
Public Survey (median per-event)	0.1451	0.2149	<0.001
Superforecasters (median per-event)	0.1110	0.3474	0.1522
Karger et al. (2024) state of the art	0.133	-	-
OpenAI o3	0.1221	0.4839	0.0025
AIA Forecaster	0.1076	-	-

Table 13: Uncertainty quantification for results on FB-7-21 in Table 2. Karger et al. (2024) do not publish individual LLM forecasts, precluding a direct pairwise comparison.

Forecaster	Brier score	Win Rate vs Best	Pairwise p-value
Public Survey (median per-event)	0.1510	0.2209	<0.001
Superforecasters (median per-event)	0.1152	0.5066	0.0946
Karger et al. (2024) state of the art	0.145	-	-
OpenAI o3	0.1262	0.4618	<0.001
AIA Forecaster	0.1099	-	-

Table 14: Uncertainty quantification for results on FB-8-14 in Table 2. Karger et al. (2024) do not publish individual LLM forecasts, precluding a direct pairwise comparison.

Forecaster	Brier score	Win Rate vs Best	Pairwise p-value
OpenAI o3	0.1324	0.7202	0.0016
AIA Forecaster	0.1258	0.7664	0.0567
Market	0.1106	-	-

Table 15: Uncertainty quantification for results on MARKETLIQUID in Table 2.

B. Evaluating Different Base Models

In this section we record the performance of different base models on the MARKETLIQUID. We choose this benchmark because it includes very recent events, all of which can be evaluated after the knowledge cutoff of the models we consider.¹⁸ We exclude the supervisor agent (see Section 7.2) to isolate the quality of the baseline reasoning, and otherwise use the pipeline described in Section 3.

Model	Brier score	Win Rate vs Best	Pairwise p-value
OpenAI GPT-4o	0.1485	0.4416	<0.001
Gemini 2.5 Pro	0.1456	0.4099	<0.001
OpenAI GPT-5	0.1412	0.5115	<0.001
Qwen-32b	0.1404	0.3155	<0.001
OpenAI o3	0.1242	0.4422	0.1561
Claude Sonnet 4	0.1195	-	-

Table 16: Performance of various base models on MARKETLIQUID

As Table 16 shows, Sonnet 4 and OpenAI o3 yield comparable performance on the MARKETLIQUID; all other models perform significantly worse.¹⁹ In addition to those presented above, we also find that OpenAI OSS-120b achieves a Brier score of 0.1076, with a win rate of 50% against Sonnet 4. However, closer inspection reveals that this

18. Claude Sonnet 4 has the latest knowledge cutoff, which is March 2025. However, its “reliable” knowledge cutoff is January 2025. The earliest questions in the MARKETLIQUID benchmark are asked “as of” 01/01/2025. This complicates the interpretation of the Sonnet 4 results, which likely include some form of lookahead bias (Sarkar and Vafa, 2024).

19. We find that Claude Opus 4 is prohibitively slow on our benchmarks and thus exclude it from our analysis.

apparently state-of-the-art performance is a mirage — the model is simply able to exploit the foreknowledge bias described in Section 5.2 more effectively than all other models we evaluate. Indeed, the analysis described in Section 5.2 reveals that 17.09% of the news queries leak foreknowledge about the event in question. These problematic queries cumulatively affect 57% of our benchmark, making it virtually impossible to evaluate the results.

We default to OpenAI o3 as our base model unless otherwise specified. Unlike Sonnet 4, its pretraining cutoff (October 2023) is well before the events which appear in FORECAST-BENCH and MARKETLIQUID, which precludes any form of pretraining lookahead bias on either benchmark (Sarkar and Vafa, 2024).

C. Sources of Foreknowledge Bias

Through extensive analysis of search API responses across SEARCH-A and SEARCH-B, we find that the primary mechanism for introducing foreknowledge bias is responses which return web pages or news articles that are updated after initial publication. They can be updated because of (1) live data feeds embedded on the page and/or (2) web pages that are continuously updated by the publisher.

Live data feeds that bypass API date filters. When search APIs crawl and summarize web pages, they often capture real-time widgets that display current data regardless of our specified date constraints. Consider this concrete example: we called the SEARCH-A API requesting information about Tapestry Inc. (TPR) stock with an explicit July 21, 2024 cutoff date. Despite this restriction, the API returned:

“As of July 23, 2025, TPR’s price was approximately \$108.72, meaning on July 21, 2024, the price was likely lower but near analyst forecasts of around \$91.82...”

The API pulled this future price from a live widget on TipRanks, defeating the purpose of our temporal restriction. This occurs because the search APIs retrieve the current version of web pages, which often embed dynamically updating content.

Search APIs return the current versions of continuously updated pages. Wikipedia and similar platforms pose a distinct challenge: search APIs typically return the latest version of these pages, not a historical snapshot, even when a past date is specified.

Search APIs returning current versions of continuously updated pages. Wikipedia and similar platforms present a different challenge: when search APIs retrieve these pages, they get the most recent version rather than a historical snapshot, even when we specify a past date.

For instance, when we use the SEARCH-A API to ask whether professional chess player Arjun Erigaisi will maintain his current ELO ranking, with an explicit July 21, 2024 cutoff date, the API returned:

“...By July 2024, he had entered the top 5 Open with a rating of 2778 after gaining 17 points. As of December 2024, he became the 16th player in history to cross the 2800 rating mark... The Tata Steel Chess Tournament 2025, where he finished 10th, might affect his standings slightly...”

Despite the intended date cutoff, the API revealed that Erigaisi crossed the 2800 threshold in December 2024, and even provided results from a January 2025 tournament. This occurred because the API pulled from a current, recently updated Wikipedia page, rather than from a historical snapshot.

D. Foreknowledge Analysis Prompt

Foreknowledge Analysis Prompt - Overview

You are analyzing a forecast question where a language model made a very confident prediction that turned out to be correct. Your task is to determine if there's evidence that the model may have used FOREKNOWLEDGE of the outcome rather than making a genuine prediction.

QUESTION: {question}

MODEL'S KNOWLEDGE CUTOFF: {news_end_date}

RESOLUTION DATE: {resolution_date}

CORRECT ANSWER: {resolution}

MODEL: {model_name}

MODEL'S PROBABILITY ASSIGNMENT: {probability:.4f} (very confident)

MODEL'S REASONING AND DATA: {model_output}

Analysis Criteria

Please analyze the model's reasoning to determine if there's evidence of foreknowledge. Look for:

1. **CLEAR EVIDENCE of outcome knowledge:** Does the reasoning explicitly reference the actual outcome, result, or resolution that occurred?
2. **Past-tense language about future events:** Does the model speak about events after the knowledge cutoff date as if they already happened? Recall that the model's knowledge cutoff is `{news_end_date}`, whereas the resolution date is `{resolution_date}`.
3. **Explicit outcome statements:** Does the model state facts about what actually occurred rather than what might occur?
4. **Access to post-event information:** Does the reasoning reference news or information that could only be available after the event resolved?

IMPORTANT: We are NOT penalizing correct predictions or good forecasting. We are only looking for cases where the model clearly had access to information about the actual outcome.

Recall that you have access to a search tool through which you can retrieve up to date information. Please feel free to use this search tool to investigate whether particular statements made by the model indicate knowledge of events after the knowledge cutoff date.

Required Response Format

Return your analysis in this JSON format:

```
{
  "has_foreknowledge": boolean (true if clear evidence of
                                foreknowledge exists),
  "confidence_level": "high" | "medium" | "low" (how confident you are
                                                in the foreknowledge assessment),
  "evidence_quotes": ["exact quote 1 showing foreknowledge",
                      "exact quote 2", ...],
  "evidence_explanation": "detailed explanation of why the evidence
                           suggests foreknowledge",
  "legitimate_reasoning": boolean (true if the reasoning could be
                                   legitimate forecasting without foreknowledge),
  "key_indicators": ["indicator 1", "indicator 2", ...] (specific signs
                                                       of foreknowledge),
  "overall_assessment": "brief summary of whether this appears to be
                         genuine forecasting or potential cheating"
}
```

Be very careful to distinguish between:

- Strong forecasting based on available information (NOT foreknowledge)
- Clear evidence of knowing the actual outcome (IS foreknowledge)

E. Further Search Details and Results

The following URLs were excluded due to frequent post-cutoff leakage via dynamic widgets or continuously updated content (see Sec. 5.2).

- https://en.wikipedia.org/wiki/FIDE_rankings
- https://en.wikipedia.org/wiki/Sarasadat_Khademalsharieh
- https://en.wikipedia.org/wiki/Nodirbek_Abdusattorov
- <https://weatherspark.com/h/m/56493/2024/7/Historical-Weather-in-July-2024-in-Strasbourg-France>
- <https://weatherspark.com/h/m/147731/2024/7/Historical-Weather-in-July-2024-at-Brest-Brittany-France>
- <https://weatherspark.com/h/m/50604/2024/7/Historical-Weather-in-July>
- <https://www.macrotrends.net/stocks/charts/HUM/humana/stock-price-history>
- <https://www.nasdaq.com/market-activity/stocks/tsco>
- https://en.wikipedia.org/wiki/Viswanathan_Anand
- <https://www.historique-meteo.net/caraibes/guadeloupe/pointe-a-pitre/2024/07/>

In Table 17, We include additional results for the LLM-as-a-judge described in Section 5.2. These estimates suggest that $\sim 1.65\%$ of search results include foreknowledge bias.

Quantity (N=4411)	Estimate	Notes
Flags (judge positive)	236	5.35% of responses
Estimated True Positives (TP)	≈ 47	$= \text{Precision} \times \text{flags}$
Estimated False Positives (FP)	≈ 189	$= \text{flags} - \text{TP}$
Estimated False Negatives (FN)	≈ 26	from $R = TP / (TP + FN)$
Estimated True Negatives (TN)	≈ 4149	$= \text{clean} - \text{FN}$
Estimated Foreknowledge Bias	$\approx 1.65\%$	$(TP + FN) / N$

Table 17: Estimating the true rate of foreknowledge bias from the 5.35% of flagged search results. Precision and recall estimates are derived from extensive manual validation of flagged responses, as described in Section 5.2

F. MarketLiquid Dataset Construction

In this section, we describe our prompts for rewording prediction market questions and our method of selecting date cutoffs. The date cutoffs depend on the length of time the market was open and were selected in consultation with a domain expert.

We first show the prompt used to reword questions:

Question rewording prompt

The following information is from a prediction market where the title is templatized and may be missing information.

The rules contain more information that should be enough to fully resolve the market.

Title: {title}

Rules: {rules}

Please output a single yes or no question that encompasses the title and rules.

When possible, keep the phrasing as close to the title as possible, but ensure the question is a natural sentence.

ONLY output the question, do not output anything else.

Next, we show the pseudocode used to select the date cutoffs:

```

open_time = ...
close_time = ...
diff = close_time - open_time
if diff.days > 91:
    days_from_close = [1, 7, 30, 60, 90]
elif diff.days > 30:
    days_from_close = [1, 7, 14, 21, diff.days]
elif diff.days > 7:
    days_from_close = [1, 3, 5, 7, diff.days]
else:
    # Invalid

```

G. LLMs Attenuate Forecasts

G.1 Examples of Hedging Behavior

Partial Reasoning Trace Showing Hedging

QUESTION: Will the total return of the ICE BofA US Corporate Index, which tracks the performance of corporate debt issued in the US domestic market, have increased by 2024-08-20 as compared to its value on 2024-07-21?

MODEL'S KNOWLEDGE CUTOFF: 2024-07-21

RESOLUTION DATE: 2024-08-20

CORRECT ANSWER: 1.0

MODEL: o3

MODEL'S PROBABILITY ASSIGNMENT: 0.6

MODEL'S REASONING AND DATA: The historical base rate of one-month positive returns for the ICE BofA US Corporate Index (or close analogues) is roughly 61%. Current carry does give a small edge, but markets already discount the possibility of a September rate cut; the asymmetric risk is that the Fed sounds ‘higher-for-longer,’ which could push yields up more than the carry buffer. Summer illiquidity and election/geopolitical headlines further skew the distribution to the downside. These considerations trim the edge implied by the initial 0.62 estimate back to an even 0.60, keeping the forecast slightly above the base rate but acknowledging that the upside case is already partly priced in.

Partial Reasoning Trace Showing Hedging

QUESTION: According to Wikipedia, will César Cielo still hold the world record for 50m freestyle in long course (50 metres) swim?

MODEL'S KNOWLEDGE CUTOFF: 2024-07-21

RESOLUTION DATE: 2024-08-20

CORRECT ANSWER: 1.0

MODEL: o3

MODEL'S PROBABILITY ASSIGNMENT: 0.97

MODEL'S REASONING AND DATA: The 14-year survival of Cielo's mark implies a weekly break rate of roughly 0.13%, suggesting >99% chance the record survives any random week. However, records are disproportionately broken at major meets and during taper periods; the Paris pool opens for training this week, creating a thin tail risk of a sanctioned time-trial or relay lead-off. We also add a small ‘Wikipedia-update’ risk. To hedge these low-but-non-zero unknowns, we shave the probability from 98% to 97%.

G.2 Platt Scaling is Generalized Log Odds Extremization

Log odds extremization as presented in Baron et al. (2014) is a special application of Platt scaling. In particular, log odds aggregation is the Platt scaling of the geometric mean of forecasts.

First we show that Equation 1 from Baron et al. (2014) is equivalent to Platt scaling. This allows us to express Platt scaling in a simpler form. That is, we show that $\frac{1}{1+\exp(-\alpha \log \frac{p}{1-p})}$ and $\frac{p^a}{p^a + (1-p)^a}$ are equivalent expressions. Starting with $\frac{p^a}{p^a + (1-p)^a}$ and $p = f(x)$ for some real valued function:

$$\frac{p^a}{p^a + (1-p)^a} = \frac{1}{1 + \left(\frac{1-p}{p}\right)^a} \quad (3)$$

$$= \frac{1}{1 + e^{\alpha \log\left(\frac{1-p}{p}\right)}} \quad (4)$$

$$= \frac{1}{1 + e^{-\alpha \log\left(\frac{p}{1-p}\right)}} \quad (5)$$

$$= \text{sigmoid}\left(\alpha \log\left(\frac{p}{1-p}\right)\right) \quad (6)$$

A more general version of Platt scaling exists, where the d parameter controls the sigmoidal curve to move left and right along the domain. Here, we show that the d parameter manifests as a $\log d$ term in the sigmoid of Platt scaling.

$$\frac{dp^a}{dp^a + (1-p)^a} = \frac{1}{1 + \frac{1}{d} \left(\frac{1-p}{p} \right)^a} \quad (7)$$

$$= \frac{1}{1 + e^{\alpha \log(\frac{1-p}{p}) - \log d}} \quad (8)$$

$$= \frac{1}{1 + e^{-\alpha \log(\frac{p}{1-p}) - \log d}} \quad (9)$$

$$= \text{sigmoid} \left(\alpha \log \left(\frac{p}{1-p} \right) + \log d \right) \quad (10)$$

With this connection to Platt scaling in mind, we show that Equation 2 is equivalent to Platt scaling of the geometric mean with a scaling parameter a and no shift parameter i.e. $d = 1$.

Beginning with log odds as prescribed in Equation 2, $\log \frac{\hat{p}}{1-\hat{p}}$ where \hat{p} is the aggregated scaled forecast, we can impute \hat{p} by applying a $\text{sigmoid}(\cdot)$. Then simplifying gives:

$$\begin{aligned} \hat{p} &= \text{sigmoid} \left(\log \frac{\hat{p}}{1-\hat{p}} \right) \\ &= \text{sigmoid} \left(a \frac{1}{n} \sum_{i=1}^n \log \frac{p_i}{1-p_i} \right) \\ &= \text{sigmoid} \left(\frac{1}{n} \sum_{i=1}^n \left(\log \frac{p_i}{1-p_i} \right)^a \right) \\ &= \frac{1}{1 + \exp \left(- \sum_{i=1}^n \left(\log \frac{p_i}{1-p_i} \right)^{a/n} \right)} \\ &= \frac{1}{1 + \frac{\prod_{i=1}^n (1-p_i)^{a/n}}{\prod_{i=1}^n p_i^{a/n}}} \\ &= \frac{1}{1 + \frac{\prod_{i=1}^n (1-p_i)^{a/n}}{\prod_{i=1}^n p_i^{a/n}}} \\ &= \frac{\prod_{i=1}^n p_i^{a/n}}{\prod_{i=1}^n p_i^{a/n} + \prod_{i=1}^n (1-p_i)^{a/n}} \\ &= \frac{\left(\prod_{i=1}^n p_i^{1/n} \right)^a}{\left(\prod_{i=1}^n p_i^{1/n} \right)^a + \left(\prod_{i=1}^n (1-p_i)^{1/n} \right)^a} \end{aligned}$$

Thus the log odds extremized score \hat{p} is a special application of Platt scaling. Specifically, it is the Platt scaling of the geometric mean of each p_i and $1 - p_i$, with Platt scaling parameter a and $d = 1$.